I. ASYMPTOTIC EQUIPARTITION (AEP)

Let $X$ be an i.i.d. random variable distributed according to $P_X$. Throughout this lecture we assume that random vector $X^n$ is an i.i.d, hence $P(x^n) = \prod_{i=1}^{n} P(x_i)$.

**Definition 1 (Typical set)** The typical set, $\mathcal{A}_{\epsilon}^{(n)}$, with respect to $P_X$, is the set of sequences $(x_1, x_2, ..., x_n) \in X^n$ with the property

$$H(X) - \epsilon \leq -\frac{1}{n} \log P(x^n) \leq H(X) + \epsilon$$

(1)

**Theorem 1 (Properties of typical set)** Let $X^n$ be an i.i.d. sequence distributed according to $P_X(x)$. For every $\epsilon > 0$ and $n$ sufficiently large, we can show that the set $\mathcal{A}_{\epsilon}^{(n)}$, has the following properties:

1) if $x^n \in \mathcal{A}_{\epsilon}^{(n)}$, then,

$$2^{-n(H(X)+\epsilon)} \leq \Pr X^n = x^n \leq 2^{-n(H(X)-\epsilon)}$$

(2)

2) For $n$ sufficiently large

$$\Pr(\mathcal{A}_{\epsilon}^{(n)}) \geq 1 - \epsilon$$

(3)

3) $|\mathcal{A}|$ denotes the number of elements (cardinality) in the set $\mathcal{A}$. The cardinality of the typical set is upper bounded by

$$|\mathcal{A}_{\epsilon}^{(n)}| \leq 2^{n(H(X)+\epsilon)}$$

(4)

4) And is lower bounded by

$$|\mathcal{A}_{\epsilon}^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$$

(5)
Informally: The typical set has probability nearly 1, all elements in it are nearly equiprobable, and the number of elements in it is nearly $2^{nH(X)}$.

Proof:
1) follows from the definition of $\mathcal{A}_\epsilon^{(n)}$.
2) follows from the law of large numbers.

$$\Pr\left\{ \left| \frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon \right\} \xrightarrow{\text{L.L.N}} 1$$  \hspace{1cm} (6)

3) Consider

$$1 = \sum_{x^n \in X^n} P(x^n) \geq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} P(x^n) \geq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = |\mathcal{A}_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)},$$  \hspace{1cm} (7)

where (a) follows from the fact that we sum over a smaller set, (b) from Eq. (2). Note that (7) implies (4).

4) According to (3) for $n$ sufficiently large we have

$$\Pr\{\mathcal{A}_\epsilon^{(n)}\} \geq 1 - \epsilon$$  \hspace{1cm} (8)

Now consider

$$1 - \epsilon \leq \Pr(\mathcal{A}_\epsilon^{(n)}) = \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = |\mathcal{A}_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}$$  \hspace{1cm} (9)

Finally, note that (9) implies (5).

II. FIXED LENGTH LOSSLESS SOURCE CODING

Recall from the examples of the introductory chapter that there are 2 modes for compression of information (source coding): lossless and lossy. Lossless compression
is usually divided into 2 types: Fixed length coding (also known as blockcoding) and variable length coding. We have seen that, for compression, it is advantageous to code in groups of symbols or blocks, even when the source is memoryless.

The conceptually simplest encoder takes a source of length $N$ and maps it to a binary word of length $NR$ so the resulting compression ratio is $\frac{\text{bits}}{\text{source symbol}}$.

![Source coding with a fixed block-length code](image)

We ask that

$$\lim_{n \to \infty} \Pr(\hat{X}^n \neq X^n) = 0$$

**Definition 2 (Fixed-length source code)** The fixed-length source code $(n, 2^{nR})$ includes:

1) Encoder $f: X^n \rightarrow \{0, 1\}^{nR}$
2) Decoder $g: \{0, 1\}^{nR} \rightarrow \hat{X}^n$

### III. Achievable Rate

**Definition 3 (An achievable rate)** A rate $R$ is achievable if there exists a sequence of codes $(n, 2^{nR})$ such that:

$$\lim_{n \to \infty} \Pr(\hat{X}^n \neq X^n) = 0$$

Let $R^*$ denote the infimum over all achievable rates. The next theorem relates $R^*$ which is defined operationally to a mathematical quantity.

**Theorem 2 (Lower bound of $R$)** For a memoryless source the smallest achievable rate $R^*$ satisfies

$$R^* = H(X)$$

The proof includes two parts: achievability and converse. In the achievability part we need to show that if $R > H(X)$, then there exists a sequence of fixed-length source codes
(n,2^{nR}) such that \( \Pr\{X^n = \hat{X}^n\} = 1 \). In the converse part we need to show that given that \( R \) is achievable, i.e., there exists a sequence of codes \((n,2^{nR})\) s.t. \( \Pr\{X^n = \hat{X}^n\} = 1 \), then \( R > H(X) \).

**Proof:** Proof of achievability part: We need to show that if \( R > R^* \), then there exists a sequence of codes \((n,2^{nR})\) such that \( P(\hat{X}^n = X^n) \rightarrow 1 \).

Let's fix \( R \), such that \( R = H(X) + \epsilon \).

**Code Design:** Assign to each sequence in \( A^{(n)}_\epsilon \) an index.

**Encoder:** If \( x^n \in A^{(n)}_\epsilon \) then sends the index \( \{1,...,2^{nR}\} \) and if not, sends 0000...0.

**Decoder:** Looks at the index, and constructs \( \hat{x}^n \) that corresponds to the index.

**Error analysis:** We have an error if the source \( x^n \) is not in the typical set \( A^{(n)}_\epsilon \), i.e.,

\[
\Pr(\hat{X}^n \neq X^n) = \Pr(X^n \notin A^{(n)}_\epsilon)
\]

and by the law of large number this probability goes to zero for any \( \epsilon > 0 \).

In the converse we need to show that if \( R \) is achievable then \( R \geq H(X) \). For this we will use a new inequality, called Fano’s inequality.

Suppose that we wish to estimate a random variable \( X \), by an estimator \( \hat{X} \). Further more assume that \( \Pr(\hat{X} \neq X) = \epsilon \). What can we say about \( H(X | \hat{X}) \). Intuitively, if \( \epsilon \) is very small then \( H(X | \hat{X}) \) should also be very small. The next theorem, called Fano’s inequality quantifies this intuition.

**Theorem 3 (Fano’s inequality)** For any estimator \( \hat{X} \) with \( \Pr(\hat{X} \neq X) = \epsilon \), we have

\[
H(X | \hat{X}) \leq 1 + \epsilon \log |\mathcal{X}|
\]

**Proof:** Let’s define:

\[
\lambda = \begin{cases} 
1 & \text{if } X = \hat{X} \\
0 & \text{if } X \neq \hat{X}
\end{cases}
\]

The PMF of \( \lambda \) is defined by:

\[
P(\lambda = 1) = 1 - \epsilon \\
P(\lambda = 0) = \epsilon
\]
Now we will find that the conditional entropy is:

\[
H(X|\hat{X}) \overset{(a)}{=} H(X, \lambda|\hat{X}) = H(\lambda|\hat{X}) + H(X|\lambda, \hat{X}) \\
\overset{(b)}{=} H(\lambda|\hat{X}) + P(\lambda = 0) \cdot H(X|\lambda = 0, \hat{X}) + P(\lambda = 1) \cdot H(X|\lambda = 1, \hat{X}) \\
\overset{(c)}{=} 1 + P(\lambda = 0) \cdot H(X|\lambda = 0, \hat{X}) \\
\overset{(d)}{\leq} 1 + \epsilon \log(|\mathcal{X}|)
\]

Where:

(a) follows from the fact that \(H(X, \lambda|\hat{X}) = H(X|\hat{X}) + H(\lambda|\hat{X}, X)\) where the last entropy equals zero.

(b) follows from \(H(\lambda|\hat{X}) \leq H(\lambda) \leq \log |\lambda|\).

(c) follows from the fact that if \(\lambda = 1\) then \(X = \hat{X}\) and then \(H(X|\lambda = 1, \hat{X}) = 0\).

(d) follows from the fact that \(H(X|\lambda = 0, \hat{X})\) is bounded by \(\log |\mathcal{X}|\) and \(P(\lambda = 0) = P(X \neq \hat{X})\).

Proof:

**Proof of the Converse part of Theorem 2:** Fix a code with a rate \(R\), i.e., \((n, 2^{nR})\) with a probability of error \(P_e^n = \Pr(X^n \neq \hat{X}^n)\).

Let’s define \(T = f(X^n) \in \{1, ..., 2^{nR}\}\), from this definition we get \(nR \geq H(T)\).

Consider\(^1\)

\[
nR \geq H(T) = I(X^n; T) = H(X^n) - H(X^n|T) = H(X^n) - H(X^n|T, \hat{X}^n)
\]

\(^1\)as an exercise, please explain each step of (15)
\[ \geq nH(X) - H(X^n | \hat{X}^n) \]
\[ \geq nH(X) - 1 - P_e^n n \log(|\mathcal{X}|) \]  
(15)

Hence we obtained that

\[ R \geq H(X) - \frac{1}{n} - P_e^n \log(|\mathcal{X}|) \]  
(16)

Now, since \( R \) is achievable there exists a sequence of codes \((n, 2^n R)\) where \( \lim_{n \to \infty} P_e^n \to 0 \). This means that for \( n \) large enough \( P_e^n \) is arbitrary small (and positive) therefore (16) implies that \( R \geq H(X) \).

\section*{A. Lossless source coding with side information}

In many distributed applications, the receiver may have some prior side information about \( X \), before it is sent. Source coding with side information addresses encoding schemes that exploit the side information in order to reduce the length of the code.

In this case assume \((X, Y)\) are i.i.d with \( P(x, y) \). A code \((n, 2^n R)\) is defined as

\begin{definition}[Fixed-length source code with side information]

The fixed-length source code \((n, 2^n R)\) includes:

1) Encoder \( f : \mathcal{X}^n \times \mathcal{Y}^n \to \{0, 1\}^{nR} \)
2) Decoder \( g : \{0, 1\}^{nR} \times \mathcal{Y}^n \to \hat{X}^n \)

An achievable rate is defined as in the case that there is no side information, namely Def. 3.

Let us denote \( R^*_{X|Y} \) the infimum achievable rate.

\begin{theorem}[Infimum for lossless reconstruction]

The infimum achievable rate for lossless reconstruction of \( X \) where side information \( Y \) is available at the encoder and decoder is given

\[ R^*_{X|Y} = H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) \]  
(17)
Proof: Achievability: If $R > R^*$, then exists a sequence of codes such that $\Pr(\hat{X}_n \neq X^n) \to 0$.

Let’s fix $R$, such that $R = H(X|Y) + \epsilon$.

Code design: For each sequence in $X^n$, we have sequence of $Y^n$ of the same length. Each time we have in $Y = 1$, we take the corresponding bits of $X$ and construct a code. The corresponding bits of $X^n$ will be a sequence in $A_{0e}^{(n)}$ or $A_{1e}^{(n)}$. Namely we assign two indexes:

1. $\{1, \ldots, 2^{n_0 R_0}\}$ for $A_{0e}^{(n)}$ of $X^n$ when $Y = 1$.
2. $\{1, \ldots, 2^{n_1 R_1}\}$ for $A_{1e}^{(n)}$ of $X^n$ when $Y = 0$.

Encoder - Consists two encoders followed by mux.

If $Y = 1$ and $x^n \in A_{0e}^{(n)}$ then send the index $\{1, \ldots, 2^{n_0 R_0}\}$ and if not, send 0000...0.
If $Y = 0$ and $x^n \in A_{t\epsilon}^{(n)}$ then send the index $\{1, ..., 2^{n_1 R_1}\}$ and if not, send 0000...0.

**Decoder** - Consist demux followed by two decoders. Each decoder looks at the index, and constructs $\hat{X}_i^n$ that corresponds to the index. Then by the value of $Y$ the demux choose the proper sequence.

And we get that:

\[ n_0 + n_1 = n \tag{18} \]
\[ nR = n_0 R_0 + n_1 R_1 \tag{19} \]

Now we can write:

\[
R = \frac{n_0}{n} R_0 + \frac{n_1}{n} R_1 \\
\geq P(y = 0) H(X|y = 0) + P(y = 1) H(X|y = 1) \\
= H(X|Y) \tag{20}
\]

And finally:

\[
\Pr(\hat{X}^n \neq X^n) \xrightarrow{n \to \infty} 0 \tag{21}
\]

thus $R \geq H(X|Y)$.

**Converse**: for the converse part, fix a scheme of rate $R$ for a block of length $n$ with a probability of error $\Pr(\hat{X}^n \neq X^n) = P_e^{(n)}$ and let $T \triangleq f(X^n, Y^n)$, consider:
\[ nR \geq H(T) \quad \text{(22)} \]
\[ \geq H(T|Y^n) \quad \text{(23)} \]
\[ \geq I(X^n; T|Y^n) \quad \text{(24)} \]
\[ \equiv H(X^n|Y^n) - H(X^n|T, Y^n) \quad \text{(25)} \]
\[ \equiv nH(X|Y) - \epsilon_n \quad \text{(26)} \]

where \( \epsilon_n \to 0 \).

(a) follows from the fact that the conditioning decreases entropy.

(b) follows from the definition of mutual information.

(c) follows Fano’s inequality and the fact that \((X_i, Y_i)\) are i.i.d.

The converse proof is completed by invoking the fact that since \( R \) is an achievable rate there exists a sequence of codes at rate \( R \) such that \( \epsilon_n \to 0 \).

\[ \blacksquare \]