Introduction to Information and Coding Theory

Lecture 5

Lecturer: Haim Permuter, Scribe: Boris Bakshan, Ronen Peker and Yaniv Nissenboim

I. Asymptotic Equipartition (AEP)

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

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

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

A more precise notation of the typical set would be $\mathcal{A}^{(n)}(X)$ or $\mathcal{A}^{(n)}(P_X)$, since the typical set is defined by the distribution of the r.v. X. However, throughout this lecture we talk only about typical set defined by P_X and therefore we allows us to omit it[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, the set $\mathcal{A}_\epsilon^{(n)}$ has the following properties:

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

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

2) For n sufficiently large

$$\Pr(X^n \in \mathcal{A}_\epsilon^{(n)}) \geq 1 - \epsilon. \tag{3}$$

[1]Later on, in the lectures on channel capacity we will define a joint typical set where we will use the notation $\mathcal{A}^{(n)}(X, Y)$ which is the typical set of the joint (X, Y)

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

$$\left|\mathcal{A}_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)} \tag{4}$$

4) The cardinality of the typical set is lower bounded by

$$\left|\mathcal{A}_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)} \tag{5}$$

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.

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{1}{n}\log P(x^n) - H(X)\right| < \epsilon\right\} = 1 \tag{6}$$

Hence, for any $\delta > 0$ including $\delta = \epsilon$ we can find an $N(\delta)$ s.t. for all $n > N(\delta)$

$$\Pr\left\{\left|\frac{1}{n}\log P(x^n) - H(X)\right| < \epsilon\right\} \geq 1-\delta. \tag{7}$$

3) Consider

$$\begin{aligned} 1 &= \sum_{x^n \in X^n} P(x^n) \\ &\overset{(a)}{\geq} \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} P(x^n) \\ &\overset{(b)}{\geq} \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= \left|\mathcal{A}_\epsilon^{(n)}\right| 2^{-n(H(X)+\epsilon)}, \end{aligned} \tag{8}$$

where (a) follows from the fact that we sum over a smaller set, (b) from Eq. (2). Note that (8) implies (4).

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

$$\Pr\{\mathcal{A}_\epsilon^{(n)}\} \geq 1-\epsilon \tag{9}$$

Now consider

$$1-\epsilon \leq \Pr(\mathcal{A}_\epsilon^{(n)}) \leq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = \left|\mathcal{A}_\epsilon^{(n)}\right| 2^{-n(H(X)-\epsilon)} \tag{10}$$

Finally, note that (10) implies (5).

■

II. Fixed length lossless source coding

In this section we consider lossless source coding (or more precisely near lossles source coding since we ask for the error to be arbitrary small) and fixed-block length. Fixed blocklength means that we map N source symbols into NR bits. The length of the block is N. The rate is defined as the number of bits per symbol. Namely,

$R = \frac{NR}{N} \quad \frac{bits}{source\ symbol}$.

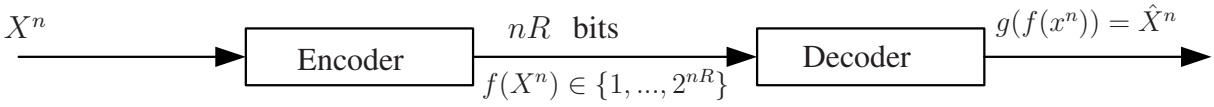

Abbildung 1. Source coding with a fixed block-length code

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

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

For a given code with rate R and blocklength n the probability of error is defined as

$$P_e^{(n)} \triangleq \Pr(\hat{X}^n \neq X^n) \tag{11}$$

or in the notation of the code

$$P_e^{(n)} \triangleq \Pr(g(f(X^n)) \neq X^n). \tag{12}$$

Figure 1 depicts the fixed-length coding system. The goal is to find a sequence of coders such that $\lim_{n\to\infty} P_e^{(n)} = 0$.

III. Highest Achievable Rate

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

$$\lim_{n\to\infty} \Pr(\hat{X}^n \neq X^n) = 0 \tag{13}$$

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

$$R^* = H(X). \tag{14}$$

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 an R that 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 of achievability part: We need to show that if $R > R^*$, then there exists a sequence of codes $(n, 2^{nR})$ such that $P(\hat{X}^n = X^n) \longrightarrow 1$. Lets fix R, such that $R = H(x) + \epsilon$. Now let's define the code and do the error analysis:

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

Encoder: If $x^n \in \mathcal{A}_\epsilon^{(n)}$ 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 $\mathcal{A}_\epsilon^{(n)}$, i.e.,

$$\Pr(\hat{X}^n \neq X^n) = \Pr(X^n \notin \mathcal{A}_\epsilon^{(n)}) \tag{15}$$

and the second property of typical sets given in Theorem 1, 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 ϵ 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

$$H(X|\hat{X}) \leq 1 + \epsilon \log |\mathcal{X}| . \tag{16}$$

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 λ is defined by:

$$P(\lambda = 1) = \bar{\epsilon},$$

$$P(\lambda = 0) = \epsilon.$$

Now we consider the conditional entropy:

$$\begin{aligned} H(X|\hat{X}) &\overset{(a)}{=} H(X, \lambda|\hat{X}) \\ &= H(\lambda|\hat{X}) + H(X|\lambda, \hat{X}) \\ &= H(\lambda|\hat{X}) + P(\lambda = 0) \cdot H(X|\lambda = 0, \hat{X}) + P(\lambda = 1) \cdot H(X|\lambda = 1, \hat{X}) \\ &\overset{(b)}{\leq} 1 + P(\lambda = 0) \cdot H(X|\lambda = 0, \hat{X}) + P(\lambda = 1) \cdot H(X|\lambda = 1, \hat{X}) \\ &\overset{(c)}{=} 1 + P(\lambda = 0) \cdot H(X|\lambda = 0, \hat{X}) \\ &\overset{(d)}{\leq} 1 + \epsilon \log(|\mathcal{X}|) \end{aligned}$$

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 than $H(X|\lambda = 1, \hat{X}) = 0$.

(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 of the Converse part of Theorem 2: Fix a code with a rate R, i.e., $(n, 2^{nR})$ with a probability of error $P_e^n = \Pr\left(X^n \neq \hat{X}^n\right)$. Let's define $T = f(\mathbf{X}^n) \in \{1, ..., 2^{nR}\}$, from this definition we get $nR \geq H(T)$. Consider[2],

$$\begin{aligned} nR &\geq H(T) \\ &= I(X^n; T) \\ &= H(X^n) - H(X^n|T) \\ &= H(X^n) - H(X^n|T, \hat{X}^n) \\ &\geq nH(X) - H(X^n|\hat{X}^n) \\ &\geq nH(X) - 1 - P_e^n n \log(|\mathcal{X}|) \end{aligned} \tag{17}$$

Hence we obtained that

$$R \geq H(X) - \frac{1}{n} - P_e^n \log(|\mathcal{X}|). \tag{18}$$

Now, since R is achievable there exists a sequence of codes $(n, 2^{nR})$ 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 (18) implies that $R \geq H(X)$. ■

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 $(n, 2^{nR})$ is defined as

Definition 4 (Fixed-length source code with side infomation) The *fixed-length source code* $(n, 2^{nR})$ includes:

1) Encoder $f : \mathcal{X}^n \times \mathcal{Y}^n \longrightarrow \{0, 1\}^{nR}$

[2]as an exercise, please explain each step of (17)

2) Decoder $g : \{0,1\}^{nR} \times \mathcal{Y}^n \longrightarrow \hat{\mathcal{X}}^n$

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

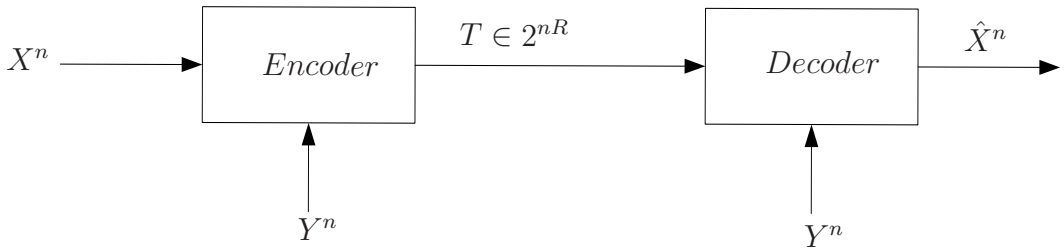

Abbildung 2. Coding with side information

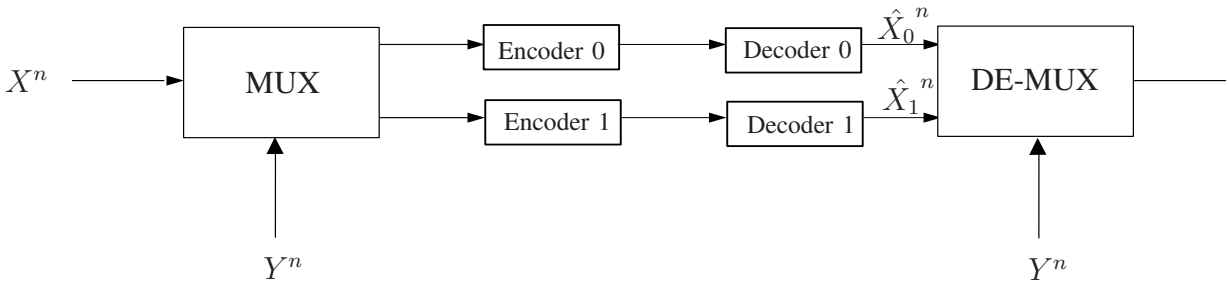

Abbildung 3. An equivalent diagram using Multiplexer

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

Theorem 4 (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

$$R^*_{X|Y} = H(X|Y) = \sum_{y \in \mathcal{X}} p(y) H(X|Y = y) \tag{19}$$

Proof: **Achievability:** If $R > R^*$, then exists a sequence of codes such that $\Pr(\hat{X}^n \neq X^n) \longrightarrow 0$.

Lets 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 $\mathcal{A}_{0\epsilon}^{(n)}$ or $\mathcal{A}_{1\epsilon}^{(n)}$. Namely we assign two indexes:

1. $\{1,...,2^{n_0 R_0}\}$ for $\mathcal{A}_{0\epsilon}^{(n)}$ of X^n when $Y = 1$.
2. $\{1,...,2^{n_1 R_1}\}$ for $\mathcal{A}_{1\epsilon}^{(n)}$ of X^n when $Y = 0$.

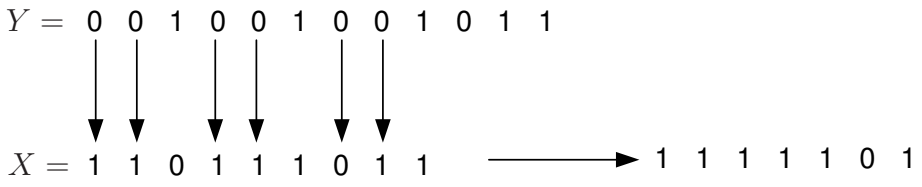

Abbildung 4. Using Y as side information source. The sequence on the right corresponds to $Y = 0$.

Encoder - Consists two encoders followed by mux.
If $Y = 1$ and $x^n \in \mathcal{A}_{0\epsilon}^{(n)}$ then send the index $\{1,...,2^{n_0 R_0}\}$ and if not, send 0000...0.
If $Y = 0$ and $x^n \in \mathcal{A}_{1\epsilon}^{(n)}$ 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:

$$n_0 + n_1 = n \tag{20}$$

$$nR = n_0 R_0 + n_1 R_1 \tag{21}$$

Now we can write:

$$\begin{aligned} R &= \frac{n_0}{n} R_0 + \frac{n_1}{n} R_1 \\ &\geq P(y=0)H(X|y=0) + P(y=1)H(X|y=1) \end{aligned}$$

$$= \quad H(X|Y) \tag{22}$$

And finaly:

$$\Pr(\hat{X}^n \neq X^n) \underset{n \to \infty}{\longrightarrow} 0 \tag{23}$$

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:

$$\begin{aligned}
nR &\geq H(T) && (24)\\
&\overset{(a)}{\geq} H(T|Y^n) && (25)\\
&\geq I(X^n; T|Y^n) && (26)\\
&\overset{(b)}{=} H(X^n|Y^n) - H(X^n|T, Y^n) && (27)\\
&\overset{(c)}{=} nH(X|Y) - \epsilon_n && (28)
\end{aligned}$$

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 $\epsilon_n \to 0$.

■