SAPIENZA
UNIVERSITÀ DI ROMA

Communications Theory and Engineering

Master's Degree in Electronic Engineering

Sapienza University of Rome

A.A. 2019-2020

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

Asymptotic Equipartition Property

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP).

The law of large numbers states that for independent, identically distributed (i.i.d.) random variables one has:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow E\,X$$

The law of large numbers

Let $X_1,\ldots,X_n$ be a sequence of i.i.d. random variables, each with mean $E(X_i) = \mu$ and standard deviation σ, we define:

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

The Weak Law of Large Numbers (WLLN) states for all $\varepsilon > 0$ then,

$$\lim_{n \to \infty} \Pr\{ | \bar{X}_n - \mu | > \varepsilon \} = 0$$

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

The asymptotic equipartition property (AEP) indicates that for a sequence of independent and identically distributed (i.i.d) random variables with probability $p(X_1, X_2, \ldots X_n)$ which is the probability of observing the sequence $X_1, X_2, \ldots X_n$ one has:

$$\frac{1}{n}\log\frac{1}{p(X_1,X_2,..,X_n)} \to H(X)$$

Proof: Functions of independent random variables are also independent random variables. Thus, since the X_i are i.i.d., so are $\log p(X_i)$. Hence, by the weak law of large numbers:

$$\frac{1}{n}\log\frac{1}{p(X_1,X_2,\ldots,X_n)} = \frac{1}{n}\sum_i \log\frac{1}{p(X_i)} \to E\log\frac{1}{p(X)} = H(X)$$

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

$$\frac{1}{n}\log\frac{1}{p(X_1,X_2,\ldots,X_n)} \to H(X)$$

This means that the probability $p(X_1,X_2,\ldots,X_n)$ assigned to an observed sequence will be close to:

$$2^{-nH}$$

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

This property allows you to divide the set of all the possible sequences into two sets:

The typical set, where the sampled entropy is close to the true entropy

The nontypical set that contains all the other sequences

Most of our attention will be on the typical sequences

SAPIENZA
UNIVERSITÀ DI ROMA

AEP: example

Suppose that a random variable $X \in \{0,1\}$ has a probability mass function defined by $p(1)=p$ and $p(0)=q$.

If $X_1, X_2, \ldots, X_n$ are i.i.d. according to $p(x)$, the probability of a sequence $X_1, X_2, \ldots, X_n$ is:

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

For example, the probability of the sequence $(1,0,1,1,0,1)$ is:

$$p^{\sum X_i} q^{n-\sum X_i} = p^4 q^2$$

Clearly, it is not true that all 2^n sequences of length n have the same probability.

SAPIENZA
UNIVERSITÀ DI ROMA

AEP

One might be able, however, to predict the probability of the sequence that is actually observed

The question is: what is the probability $p(X_1,X_2, \ldots ,X_n)$ of the outcomes $X_1,X_2, \ldots , X_n$, where $X_1,X_2, \ldots , X_n$ are i.i.d. according to $p(x)$ $(i.i.d. \sim p(x))$

The answer is: $p(X_1, X_2, \ldots X_n)$ is 2^{-nH} with high probability

In other words:

$$\Pr\left\{(X_1,X_2,\ldots,X_n): p(X_1,X_2,\ldots,X_n)=2^{-n(H\pm\varepsilon)}\right\}\approx 1$$

convergence in probability

SAPIENZA
UNIVERSITÀ DI ROMA

AEP: example

Previous example $p(1)=p$ and $p(0)=q$

If $X_1,X_2,\ldots,X_n$ are i.i.d. according to $p(x)$, the probability of the sequence $X_1,X_2,\ldots,X_n$ is:

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

According to the asymptotic equipartition property the number of "1" in the sequence is with high probability equal to np and all the sequences of this type have the same probability $2^{-nH(p)}$

Convergence of random variables

Definition: Given a sequence of random variables, $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$, converges to a random variable X:

In probability if

$$\forall \varepsilon > 0, \Pr\left\{\left|X_n - X\right| > \varepsilon\right\} \to 0$$

That is, the variables differ only for events with zero probability.

In mean square if

$$E\left(X_n - X\right)^2 \to 0$$

With probability 1 (also called almost surely) if

$$\Pr\left\{\lim_{n\to\infty} X_n = X\right\} = 1$$

SAPIENZA UNIVERSITÀ DI ROMA

Theorem AEP

If $X_1, X_2, \ldots, X_n$ are i.i.d. with $X \sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \quad \textit{in probability}$$

If $X_1, X_2, \ldots, X_n$ are independent then also $f(X_1)$, $f(X_2)$, ..., $f(X_n)$ are independent (functions of independent r.v. are also independent), that is if $X_1, X_2, \ldots, X_n$ are i.i.d. then so are $\log p(X_i)$

Hence, by the weak law of large numbers:

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_i \log p\left(X_i\right) \to E \log p(X) \quad \textit{Convergence in probability}$$
$$= H(X)$$

■

SAPIENZA
UNIVERSITÀ DI ROMA

The typical set $A_\varepsilon^{(n)}$

Definition: The typical set with respect to $p(x)$ is the set of sequences $(X_1, X_2, \ldots, X_n)$ such that:

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

SAPIENZA UNIVERSITÀ DI ROMA

The typical set $A_\varepsilon^{(n)}$

The typical set has the following properties:

$$If\ (x_1, x_2, \ldots, x_n) \in A_\varepsilon^{(n)} \quad then \quad H(X) - \varepsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \varepsilon$$

$$\Pr\left\{A_\varepsilon^{(n)}\right\} > 1 - \varepsilon \quad \text{for n sufficiently large}$$

$$\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)} \quad where \left|A_\varepsilon^{(n)}\right| is\ the\ cardinality\ of\ A_\varepsilon^{(n)}$$

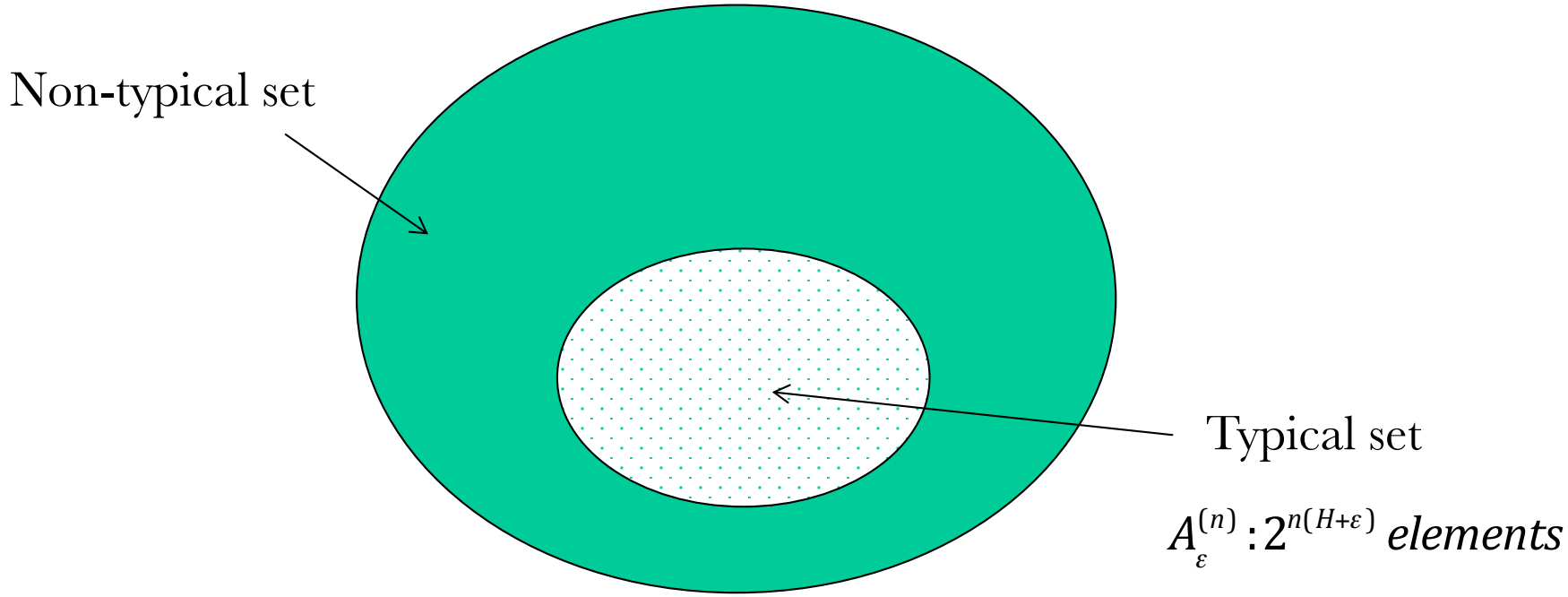
$$\left|A_\varepsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)} \quad \text{for n sufficiently large}$$

Summing up: the typical set has probability close to 1, all the elements of the typical set have same probability, and the number of elements in the typical set is close to 2^{nH}

Consequences of the AEP

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables drawn from the probability mass function $p(x)$.

We would like to find the shortest possible description of these sequences (coding)



Consequences of the AEP

Let $X^n = X_1, X_2, \ldots, X_n$ i.i.d. with pmf $p(x)$

Let $\varepsilon > 0$

there exists a code that maps sequences X^n of length n into binary strings such that the mapping is one-to-one (and therefore invertible) and

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \varepsilon \quad \text{for n sufficiently large}$$

where $l(x^n)$

indicates the length of the code word that corresponds to x^n

Therefore, it is possible to represent sequences X^n using $nH(X)$ bits on the average.

SAPIENZA
UNIVERSITÀ DI ROMA

Observation

Let $X^n = X_1, X_2, \ldots, X_n$, n be a sequence of Bernouilli (*) r.v. with parameter $p = 0.9$

The typical sequences are those that have 90% of "1" bits

However this does not include the most probable sequence composed of all "1"

So the typical set does not include the most probable sequence composed of all "1"

It can be shown that the ”high probability set” composed of the most probable sequences, that is including the typical set AND the sequence of all “1”, and the typical set have "almost" the same size

(*) *a Bernouilli r.v. (p) is a binary r.v. that takes the value 1 with probability p*

SAPIENZA
UNIVERSITÀ DI ROMA

Observation (continued)

- The concept of a typical set is generally different from that of a high probability set
- The typical set contains the "typical" sequences, i.e. those that reflect the characteristics of the source
- The high probability set contains sequences having, by their nature, high probability
- The two sets have almost the same size

SAPIENZA UNIVERSITÀ DI ROMA

Source coding theorem

Consider a source consisting of repeated tests of a r.v. X with $p(x)$ and with r trials/second

The rate R of the source in bits/s is defined as:

$$R = rH(X)$$

Theorem : the source can be coded with a source encoder in a bit stream with a transmission rate equal to R + ε, for any ε> 0

Therefore, the problem is to find codes that make ε small:
we often choose to adopt sub-optimal codes that are more easily realized

SAPIENZA UNIVERSITÀ DI ROMA

Source encoding

What are the necessary and desirable properties for source coding?

Required: a code must be uniquely decodable (there cannot be two symbols identified by the same codeword)

Desirable: the average length of a codeword should be minimized in order to minimize the bit rate required to transfer the source symbols

SAPIENZA
UNIVERSITÀ DI ROMA

Source encoding

Consider a source that emits symbols in the alphabet:

$$S = \{a_1, a_2, ..., a_K\}$$

characterized by the following probabilities:

$$p(a_1), p(a_2), ..., p(a_K)$$

and suppose we have chosen a code that associates to each of the K symbols, a codeword that has length:

$$\{l_1, l_2, ..., l_K\}$$

SAPIENZA UNIVERSITÀ DI ROMA

Code properties

The average length of a codeword is therefore given by:

$$\overline{L} = \sum_{k=1}^{K} p(a_k) l_k$$

and the code can be characterized by an efficiency defined as:

$$\eta = \frac{L_{\min}}{\overline{L}}$$

where L_{min} is the minimum average length of a codeword defined by the source coding theorem, i.e.:

$$L_{\min} = H(p)$$

SAPIENZA UNIVERSITÀ DI ROMA

Prefix rule

Once source symbols are coded into a sequence of bits, one must ensure that each symbol in the sequence is identifiable

If one could one would insert a special symbol between codewords, such as a ",". In this way each symbol in the sequence would be identifiable

However inserting a "," is not feasible in **digital** communications

The way out is codes that check the prefix rule

A code that checks the prefix rule is a variable-length code in which no codeword can be the beginning (prefix) of another codeword

SAPIENZA UNIVERSITÀ DI ROMA

Prefix rule

Example:

Symbol	Prob.of Occurrence	Code I	Code II	Code III
s_0	0.5	0	0	0
s_1	0.25	1	10	01
s_2	0.125	00	110	011
s_3	0.125	11	111	0111

Shannon-Fano coding

The Shannon-Fano code verifies the prefix rule

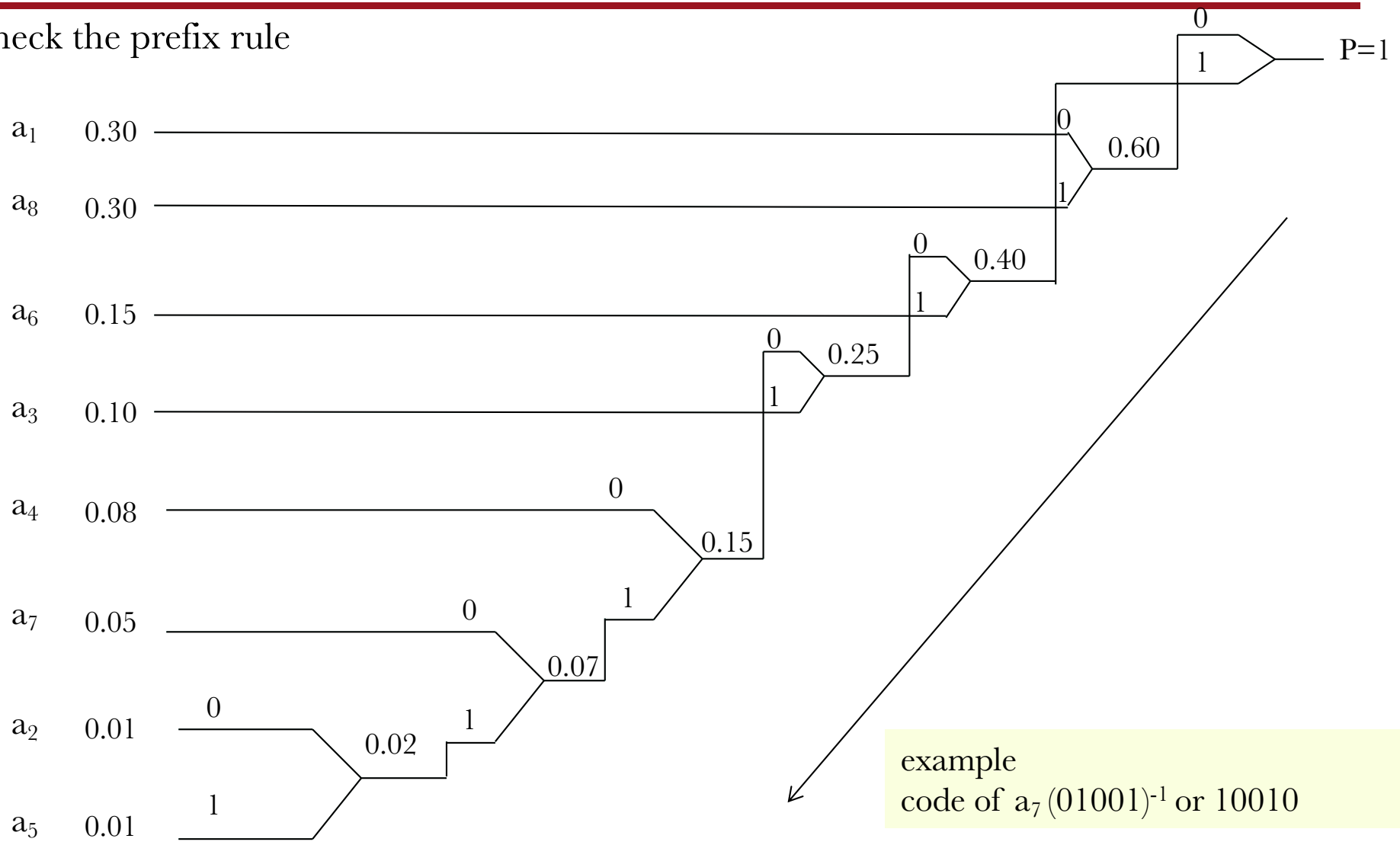
It tries to create classes that are equally likely

	symbol	probability	1° digit	2° digit	3° digit	4° digit	5° digit
0.3	a_1	0.3	0	0			
0.3	a_8	0.3	0	1			
0.6							
0.4	a_6	0.15	1	0	0		
	a_3	0.10	1	0	1		
0.25							
0.15	a_4	0.08	1	1	0		
	a_7	0.05	1	1	1	0	
	a_2	0.01	1	1	1	1	0
	a_5	0.01	1	1	1	1	1

The coding is not unique because it is not always possible to create classes that are equally likely

Huffman coding

Check the prefix rule



example
code of a_7 $(01001)^{-1}$ or 10010