Lecture 5 : typical sequences and the AEP

February 7 2020

Outline

1. Stationary source, entropy per letter ;
2. Markov source ;
3. typical sequences, AEP ;
4. Shannon's theorem for the sources satisfying the AEP ;
5. AEP for memoryless sources ;
6. AEP for stationary Markov sources.

1. Stationary Source

A source $X_1, X_2, \ldots,$ produces a sequence of letters in $\mathcal{X}$.

These random variables are not necessarily identically distributed.

These random variables are not necessarily independent. Characterized by

$$\mathbf{P}\left\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\right\} = p(x_1, \ldots, x_n), \quad n = 1, 2 \ldots$$

Entropy rate

Reminder : The entropy of the L first letters is given by

$$H(X_1, \ldots, X_L) = \sum_{x_1, \ldots, x_L} -p(x_1, \ldots, x_L) \log_2 p(x_1, \ldots, x_L)$$

The *entropy per source symbol* of a source $\mathcal{X}$ is defined by

$$H(\mathcal{X}) = \lim_{L \to \infty} \frac{1}{L} H(X_1, \ldots, X_L),$$

if this limit exists. It is also called the entropy rate.

Entropy rate of a stationary process

Definition A source is called *stationary* if its behavior does not change with a time shift, that is for all nonnegative integers n and t and all $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = p_{X_{1+t} \ldots X_{n+t}}(x_1, \ldots, x_n)$$

Theorem 1. *For any stationary source, the following limits exist and are equal*

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n \to \infty} H(X_n \mid X_1, \ldots, X_{n-1}).$$

Proof

1. Let us first show that $\lim_{L\to\infty} H(X_L \mid X_1, \ldots, X_{L-1})$ exists for a stationary source.

   For all L, we have

$$\begin{aligned} H(X_L \mid X_1, \ldots, X_{L-1}) &\leq H(X_L \mid X_2, \ldots, X_{L-1}) \\ &= H(X_{L-1} \mid X_1, \ldots, X_{L-2}) \end{aligned}$$

   Thus the sequence $(H(X_i \mid X_1, \ldots, X_{i-1}))_{i>0}$ is non increasing.

   This sequence is nonnegative $\Rightarrow$ it converges

Proof

2. For all $L > 0$, we have

$$H(X_1, \ldots, X_L) = \sum_{i=1}^{L} H(X_i \mid X_1, \ldots, X_{i-1}).$$

Therefore

$$\frac{1}{L} H(X_1, \ldots, X_L)$$

which is the average of $H(X_i \mid X_1, \ldots, X_{i-1})$, which is known to converge, also converges.

Cesaro's theorem

if a sequence $u_n \to \ell$, then $1/n \sum_{i=1}^{n} u_i \to \ell$.

Source where the entropy rate is undefined

$X_1, \ldots . X_n$ independent with $p_i = p(X_i) = 1$ such that

$$p_i = \begin{cases} 1/2 & 2^{2^{2k}} < i \leq 2^{2^{2k+1}} \\ 0 & 2^{2^{2k+1}} < i \leq 2^{2^{2k+2}} \end{cases}$$

On intervals of exponential size $H(X_i) = 1$, followed by even larger intervals of exponential size such that $H(X_i) = 0$.

We have $\frac{1}{n}H(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i)$.
Let $u_{2k+1} \stackrel{\text{def}}{=} \sum_{i\leq 2^{2^{2k+1}}} H(X_i)$ and $u_{2k} \stackrel{\text{def}}{=} \sum_{i\leq 2^{2^{2k}}} H(X_i)$ we have

$$\begin{aligned} u_{2k} - u_{2k-1} &= 0 \\ u_{2k+1} - u_{2k} &= 2^{2^{2k}}\left(2^{2^{2k}} - 1\right) \end{aligned}$$

and

$$\frac{2^{2^{2k}}\left(2^{2^{2k}} - 1\right)}{2^{2^{2k+1}}} = \frac{u_{2k+1} - u_{2k}}{2^{2^{2k+1}}} \leq \frac{u_{2k+1}}{2^{2^{2k+1}}} \leq 1$$

$$0 \leq \frac{u_{2k}}{2^{2^{2k}}} = \frac{u_{2k-1}}{2^{2^{2k}}} \leq 2^{-2^{2k-1}}$$

This averages oscillates between 0 and 1 and does not have a limit.

2. Time Invariant Markov Source

Definition A *Markov source of order 1* is such that for all nonnegative integers n and for $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$\mathbf{P}[X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}] = \mathbf{P}[X_n = x_n \mid X_{n-1} = x_{n-1}]$$

Its order is s if

$$\mathbf{P}[X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}] = \mathbf{P}[X_n = x_n \mid X_{n-1} = x_{n-1} \ldots X_{n-s} = x_{n-s}]$$

Definition The source is *time invariant* if these probabilities do not depend on n.

Notation $p(x_2 \mid x_1) = \mathbf{P}[X_n = x_2 \mid X_{n-1} = x_1]$.

Theorem

Theorem 2. *The entropy rate of a time invariant Markov source (of order 1) is equal to*

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_{x_1,x_2} -\lambda(x_1)p(x_2 \mid x_1)\log_2 p(x_2 \mid x_1)$$

where $\lambda(x), x \in \mathcal{X}$ *is the* stationary distribution.

Stationary Distribution

Let $\Pi = p(x_2|x_1)_{x_2,x_1}$. The probability distribution vector $V_n = (p(X_n = a_1), \ldots, p(X_n = a_k))$ satisfies

$$V_n = \Pi V_{n-1}$$

If the process admits a stationary distribution $\Lambda = (\lambda(x_1), \ldots, \lambda(x_k))$, it should satisfy

$$\Lambda = \Pi\Lambda$$

Exercise

Describe the stationary distribution(s) corresponding to

$$\Pi = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Proof

$$
\begin{aligned}
H(X) &= \lim_{L\to\infty}(H(X_L|X_{L-1}\dots X_1) \\
&= H(X_2|X_1) \\
&= -\sum_{x_2,x_1} p(x_2|x_1)\lambda(x_1)\log_2 p(x_2|x_1)
\end{aligned}
$$

The entropy of English/ Markov model

1) approximation of order 0 (all symbols are i.i.d)

$$H_0 = \log 27 \approx 4.76.$$

```
XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD
```

2) approximation of order 1 (the letters are chosen according to their frequency in English)

$$H_1 \approx 4.03$$

```
OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
ALHENHTTPA OOBTTVA NAH BRL
```

3) Approximation of order 2 : same distribution of the pairs as in English

```
ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMMY
ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
TIZIN ANDY TOBE SEACE CTISBE
```

4) Approximation of order 3 : same frequency of the triplets as in English

```
IN NO IST LAT WHEY CRATICT FROURE BERS GROCID
PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
REGOACTIONA OF CRE
```

5) Markov model of order 3 of English

$$H_{1/3} \approx 2.8$$

```
THE GENERATED JOB PRIVIDUAL BETTER TRAND THE
DIPLAYED CODE, ABOVERY UPONDULTS WELL THE
CODERST IN THESTICAL IT DO HOCK BOTH MERG.
(INSTATES CONS ERATION. NEVER ANY OF PUBLE AND TO
THEORY. EVENTIAL CALLEGAND TO ELAST BENERATED IN
 WITH PIES AS WITH THE)
```

6) Approximation of order 0 on the words

```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME
CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO
OF TO EXPERT GRAY COME TO FURNISHES THE LINE
MESSAGE HAD BE THESE
```

7) Markov model of order 1 on the words

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE
TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED
```

$$H_{\text{English}} \approx 1.34$$

Using a Markov model for arithmetic coding

Memoryless model

$$\mathbf{P}(x_1, x_2, \ldots, x_n) = \mathbf{P}(x_1)\mathbf{P}(x_2)\ldots\mathbf{P}(x_n)$$

Markov model of order 1

$$\mathbf{P}(x_1, x_2, \ldots, x_n) = \mathbf{P}(x_1)\mathbf{P}(x_2|x_1)\ldots\mathbf{P}(x_n|x_{n-1})$$

Markov model of order 2

$$\mathbf{P}(x_1, x_2, \ldots, x_n) = \mathbf{P}(x_1)\mathbf{P}(x_2|x_1)\mathbf{P}(x_3|x_1, x_2)\ldots\mathbf{P}(x_n|x_{n-2}, x_{n-1})$$

3. Typical Sequences

Consider a source $X_1, X_2, \ldots, X_n, \ldots$ over an alphabet $\mathcal{X}$. Assume that the entropy rate exists.

$$\mathcal{H} = H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

Definition Set of ε-typical sequences (of length n)

$$A_\varepsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in \mathcal{X}^n, \left| \frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - \mathcal{H} \right| \leq \varepsilon \right\}$$

Typical Sequences/most likely sequences

The typical sequences are not necessarily the most likely sequences !

Consider a binary source where the 0's are produced with probability $2/3$ and the 1's with probability $1/3$,

— The all 1 sequence is the least likely sequence and is not typical.
— The all 0 sequence is the most likely sequence and is not typical either.
— The frequency of 1's in the typical sequences is $\approx \frac{1}{3}$.

Asymptotic Equipartition Property

Definition [Asymptotic Equipartition Property]
A source verifies the AEP if :

$$\forall \varepsilon > 0, \lim_{n \to \infty} \mathbf{P}\left[A_\varepsilon^{(n)}\right] = 1.$$

Properties of typical sequences

Proposition 1. *For every source verifying the AEP*

1. $\frac{1}{n}\log_2 \frac{1}{p(x_1,\ldots,x_n)} \underset{n\to\infty}{\longrightarrow} \mathcal{H}$ *almost surely*
2. $|A_\varepsilon^{(n)}| \leq 2^{n(\mathcal{H}+\varepsilon)}$
3. *for* n *sufficiently large,* $|A_\varepsilon^{(n)}| \geq (1-\epsilon)2^{n(\mathcal{H}-\varepsilon)}$.

In other words :

there are $\approx 2^{n\mathcal{H}}$ typical sequences, each of them has probability $\approx 2^{-n\mathcal{H}}$.

Proof

1. By definition,

$$\left|\frac{1}{n}\log_2\frac{1}{p(x_1,\ldots,x_n)}-\mathcal{H}\right|\leq\varepsilon$$

with probability $\mathbf{P}[A_\varepsilon^{(n)}]$. Therefore

$$\mathcal{H}\quad-\quad\varepsilon\qquad\leq\qquad\frac{1}{n}\log_2\frac{1}{p(x_1,\ldots,x_n)}\qquad\leq\qquad\mathcal{H}\quad+\quad\varepsilon$$

with probability $\mathbf{P}[A_\varepsilon^{(n)}]$. If the source satisfies the AEP, $\mathbf{P}[A_\varepsilon^{(n)}]\to 1$.

Proof

2. and 3. Let $\varepsilon > 0$ and n be an integer. We have

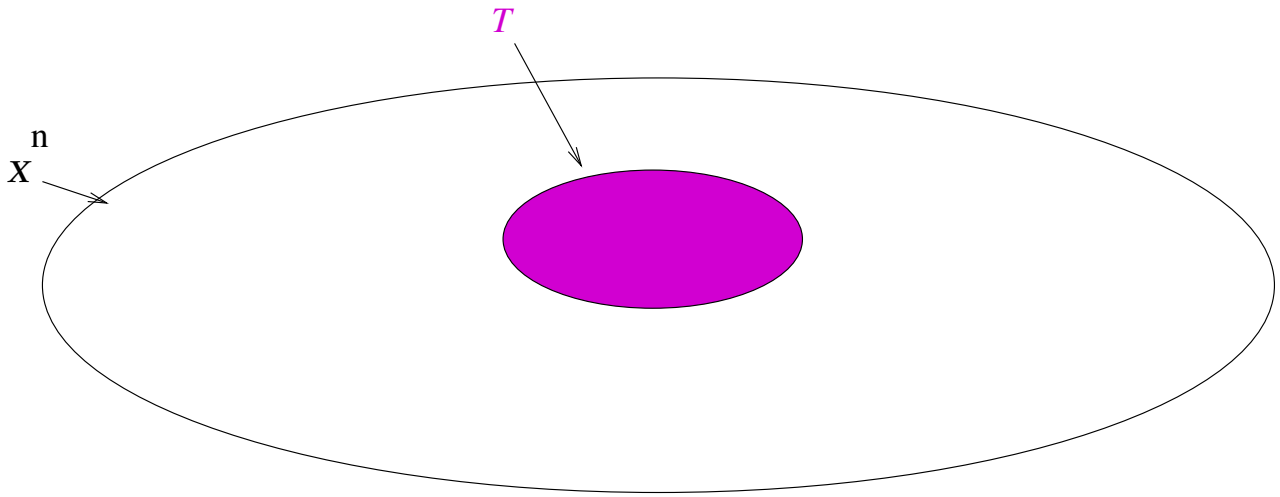
$$\begin{aligned} |A_\varepsilon^{(n)}| 2^{-n(\mathcal{H}+\varepsilon)} &\leq \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} \mathbf{P}[X_1 = x_1, \ldots, X_n = x_n] \\ &= \mathbf{P}[A_\varepsilon^{(n)}] \leq 1, \end{aligned}$$

and therefore $|A_\varepsilon^{(n)}| \leq 2^{n(\mathcal{H}+\varepsilon)}$. We also have

$$\mathbf{P}[A_\varepsilon^{(n)}] = \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} \mathbf{P}[X_1 = x_1, \ldots, X_n = x_n] \leq |A_\varepsilon^{(n)}| 2^{-n(\mathcal{H}-\varepsilon)},$$

which gives $|A_\varepsilon^{(n)}| \geq \mathbf{P}[A_\varepsilon^{(n)}] 2^{n(\mathcal{H}-\varepsilon)}$

The picture



The probability distribution is concentrated over only $2^{n(H+\epsilon)}$ $(\ll |\mathcal{X}|^n)$ sequences.

Ensembles of probability ≈ 1

Theorem 3. *Consider a source satisfying the AEP. Let* $B_\delta^{(n)}$ *s.t.*

$$\mathbf{P}\left\{B_\delta^{(n)}\right\} \geq 1-\delta$$

then

1. *for all* $\delta' > 0$,

$$\frac{1}{n}\log|B_\delta^{(n)}| > H-\delta', \text{ for } n \text{ sufficiently large.}$$

2.

$$\mathbf{P}\left\{A_\varepsilon^{(n)} \cap B_\delta^{(n)}\right\} \geq 1-\varepsilon-\delta \text{ for } n \text{ sufficiently large.}$$

Proof

Fact : if $\mathbf{P}(A) \geq 1-\varepsilon_1$ and $\mathbf{P}(B) \geq 1-\varepsilon_2$ then $\mathbf{P}(A \cap B) \geq 1-\epsilon_1-\epsilon_2$

$$\begin{aligned} 1-\varepsilon-\delta &\leq \mathbf{P}\left\{A_\varepsilon^{(n)} \cap B_\delta^{(n)}\right\} \\ &= \sum_{x^n \in A_\varepsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\ &\leq \sum_{x^n \in A_\varepsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\varepsilon)} \\ &\leq |B_\delta^{(n)}| 2^{-n(H-\varepsilon)} \end{aligned}$$

The conclusion follows by taking the $\log_2$.

4. Shannon's Theorem

Definition Let φ be a source coding for a source $\mathcal{X}$, its average length/letter is defined by

$$\mathcal{L}(\varphi) = \lim_{n\to\infty} \frac{1}{n} \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n)|\varphi(x_1,\ldots,x_n)|,$$

when this limit exists.

Theorem 4. **[Shannon]** *For a discrete source $\mathcal{X}$ of entropy rate $\mathcal{H}$ which satisfies the AEP.*

1. *Every unambiguous coding φ for $\mathcal{X}$ satisfies $\mathcal{L}(\varphi) \geq \mathcal{H}$.*
2. *there exists an unambiguous coding φ for $\mathcal{X}$ such that $\mathcal{L}(\varphi) \leq \mathcal{H} + \varepsilon$, $\forall \varepsilon > 0$.*

Proof

1. Exercise : use Theorem 2.

Proof

2. Let $\varepsilon > 0$, and $n > 0$. For every $n > 0$, let
   1. F_n be a fixed length code of minimal length for $\mathcal{X}^n$
   2. $G_{n,\varepsilon}$ be a fixed length code of minimal length for $A_\varepsilon^{(n)}$.
   3. $\varphi_{n,\varepsilon}$ a code for $\mathcal{X}^n$ defined by

$$\varphi_{n,\varepsilon}(x_1,\ldots,x_n) = \begin{cases} 0 \parallel G_{n,\varepsilon}(x_1,\ldots,x_n) & \text{if } (x_1,\ldots,x_n) \in A_\varepsilon^{(n)} \\ 1 \parallel F_n(x_1,\ldots,x_n) & \text{otherwise} \end{cases}$$

Proof (cont'd)

By choosing n sufficiently large $\mathbf{P}[A_\varepsilon^{(n)}] \geq 1 - \varepsilon$.

$$
\begin{aligned}
|\varphi_{n,\varepsilon}| &= \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)|\varphi(x^n)| + \sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)|\varphi(x^n)| \\
&\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)(n(H+\varepsilon)+2) + \sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)(n\log|\mathcal{X}|+2) \\
&= \mathbf{P}[A_\varepsilon^{(n)}](n(H+\varepsilon)+2) + \mathbf{P}[\overline{A_\varepsilon^{(n)}}](n\log|\mathcal{X}|+2) \\
&\leq n(H+\varepsilon) + \varepsilon n(\log|\mathcal{X}|) + 2
\end{aligned}
$$

The conclusion follows by taking $n \to \infty$.

5. AEP of memoryless sources

A memoryless source verifies the AEP.

Recall that :

$$\frac{1}{L}H(X_1,\ldots,X_L) = \frac{1}{L}(LH(X_1)) = H(X_1) = -\sum p(x_i)\log_2 p(x_i).$$

Memoryless source

Reminder : weak law of large numbers Let $Z_1, Z_2, \ldots, Z_n, \ldots$ be a sequence of i.i.d. random variables of expectation μ. Let

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i,$$

be the average of the first n r.v.

Then for all ε :

$$\mathbf{P}\left\{|\bar{Z}_n - \mu| > \varepsilon\right\} \to 0$$

when $n \to \infty$ (weak law of large numbers).

A theorem

Let $X_1, \ldots, X_n, \ldots$ be i.i.d variable with p.d. $p(X)$ then

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \quad = \quad -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i)$$

From the weak law of large numbers

$$\mathbf{P}\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) - \mathbb{E}\left[-\log p(X)\right]\right| \leq \varepsilon\right\} \to 1$$

Rewriting

$$\mathbf{P}\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) - \mathbb{E}\left[-\log p(X)\right]\right| \leq \varepsilon\right\} \to 1$$

$\mathbb{E}\left[-\log p(X)\right] = H$. Therefore

$$\mathbf{P}\left\{\left|-\frac{1}{n}\log p(X_n, \ldots, X_1) - H\right| \leq \varepsilon\right\} \to 1$$

for all ϵ :

$$\mathbf{P}(A_n^{(\varepsilon)}) \to 1$$

when $n \to \infty$.

Typical sequences of a discrete memoryless source

Let $x = (x_1, \ldots, x_n)$ be a sequence of length n, for a source with alphabet $\mathcal{X} = \{a_1, \ldots, a_k\}$. Let $n_{a_i}(x)$ be the number of times that a_i occurs in x, and $n_{a_i}(x)/n$ be the associated frequency.

The sequence x is ε-typical of length n iff

$$\left| \sum_{i=1}^{k} \left(\frac{n_{a_i}(x)}{n} - p(a_i) \right) \log p(a_i) \right| < \varepsilon$$

The two definitions are equivalent in the case of a discrete memoryless source.

6. Markov Source

Proposition 2. *A time invariant Markov source satisfies the AEP.*

Proof

Let $\mathcal{X} = X_1, \ldots, X_l, \ldots$. We have $H(\mathcal{X}) = H(X_2|X_1)$. We want to prove that

$$\mathbf{P}\left\{\left|-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| < \varepsilon\right\} \to 1$$

On one hand

$$\begin{aligned} -\frac{1}{n}\log_2 p(x_n, \ldots, x_1) &= -\frac{1}{n}\log_2 p(x_n|x_{n-1})\ldots p(x_2|x_1)p(x_1) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log_2 p(x_i|x_{i-1}) - \frac{\log p(x_1)/p(x_1|x_0)}{n} \end{aligned}$$

Proof (cont'd)

Consider $-\frac{1}{n}\sum_{i=1}^{n}\log_2 p(x_i|x_{i-1})$ and the random variables $Z_i = -\log_2 p(X_i|X_{i-1})$. They are i.i.d., and their average $\bar{Z}_n$ satisfies the weak law of large numbers

$$\mathbf{P}\left\{|\bar{Z}_n - \mu| < \varepsilon\right\} \to 1 \text{ where}$$

$$\mu = E(Z_i) = \sum_{x_i, x_{i-1}} -p(x_i, x_{i-1})\log_2 p(x_i|x_{i-1}) = H(X_2|X_1) = H(\mathcal{X})$$

Observe that $\bar{Z}_n = -\frac{1}{n}\sum \log_2 p(x_1, \ldots, x_n) + \frac{\log p(x_1)/p(x_0|x_1)}{n}$

We have $\mathbf{P}\left\{\left|-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| < \varepsilon\right\} \to 1.$