LUND UNIVERSITY

Information Theory

Lecture 7
AEP and its consequences

STEFAN HÖST

Law of large numbers

Theorem (The weak law of large numbers)

Let $X_1, X_2, \ldots, , X_n$ be i.i.d. random variables with expectation $E\big[X\big]$. Then, the arithmetic mean converges (in probability) to the expectation,

$$\frac{1}{n}\sum_i X_i \xrightarrow{p} E\big[X\big]$$

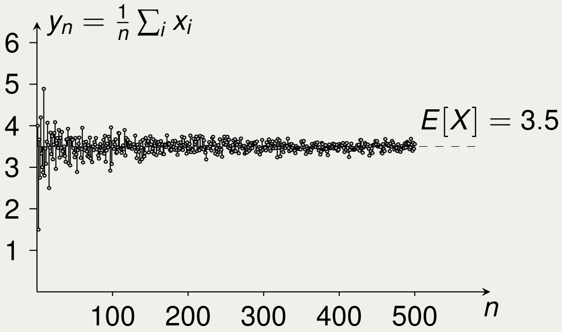
Stated differently, this means that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_i X_i - E\big[X\big]\Big| < \varepsilon\Big) = 1$$

Fair die

Example

Consider n consecutive rolls with a fair die, giving the results vector $\boldsymbol{x} = (x_1, \dots, x_n)$. Let $y_n = \frac{1}{n}\sum_i x_i$ be the average.

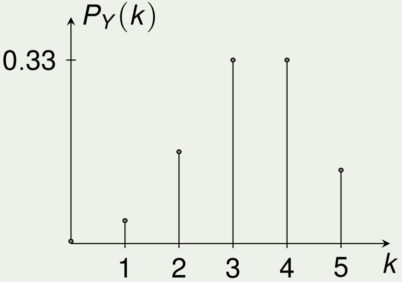


Binary vector

Example

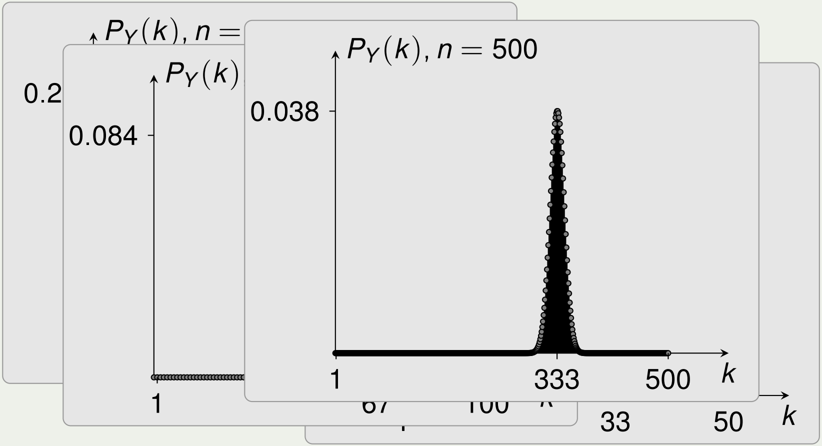
Consider a binary length 5 vector $\boldsymbol{X} = (X_1, X_2, \dots, X_5)$, where X_i i.i.d. with $p(1) = \frac{2}{3}$. Let $Y = \sum X_i$ be the number of ones,

k	$P_Y(k) = \binom{5}{k}\frac{2^k}{3^5}$
0	0.0041
1	0.0412
2	0.1646
3	0.3292
4	0.3292
5	0.1317



Binary vector

Example (cont'd)



AEP
(Asymptotic Equipartition Property)

Definition (AEP)

The set of ε-typical sequences $A_\varepsilon(X)$ is the set of all n-dimensional vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ such that

$$\left| -\frac{1}{n} \log p(\boldsymbol{x}) - H(X) \right| \leq \varepsilon$$

AEP (Alternative definition)

The ε-typical sequences can definition as the set of vectors $\boldsymbol{x}$ such that

$$2^{-n(H(X)+\varepsilon)} \leq p(\boldsymbol{x}) \leq 2^{-n(H(X)-\varepsilon)}$$

Binary vector

Example (cont'd)

Consider a binary 5-dimensional vector where $p(1) = \frac{2}{3}$. The entropy is $h(1/3) = 0.918$. Let $\varepsilon = 0.138$ (15% of $h(p)$).

$\boldsymbol{x}$	$p(\boldsymbol{x})$	
00000	0.0041	
00001	0.0082	
00010	0.0082	
00011	0.0165	
00100	0.0082	
00101	0.0165	
00110	0.0165	
00111	0.0329	$\star$
01000	0.0082	
01001	0.0165	
01010	0.0165	

$\boldsymbol{x}$	$p(\boldsymbol{x})$	
01011	0.0329	$\star$
01100	0.0165	
01101	0.0329	$\star$
01110	0.0329	$\star$
01111	0.0658	$\star$
10000	0.0082	
10001	0.0165	
10010	0.0165	
10011	0.0329	$\star$
10100	0.0165	
10101	0.0329	$\star$

$\boldsymbol{x}$	$p(\boldsymbol{x})$	
10110	0.0329	$\star$
10111	0.0658	$\star$
11000	0.0165	
11001	0.0329	$\star$
11010	0.0329	$\star$
11011	0.0658	$\star$
11100	0.0329	$\star$
11101	0.0658	$\star$
11110	0.0658	$\star$
11111	0.1317	

AEP($\star$): $0.026 \leq p(\boldsymbol{x}) \leq 0.067$

AEP

Theorem

For each ε there exists an integer n_0 such that, for each $n > n_0$, $A_\varepsilon(X)$ fulfills

1. $P\big(\boldsymbol{x} \in A_\varepsilon(X)\big) \geq 1 - \varepsilon$
2. $(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq \big|A_\varepsilon(X)\big| \leq 2^{n(H(X)+\varepsilon)}$

Example

Example (cont'd)

Let $\varepsilon = 0.046$ (5% of $h(1/3)$).

n	$(1-\varepsilon)2^{n(H(X)-\varepsilon)}$	$\leq$ $\lvert A_\varepsilon(X)\rvert$ $\leq$	$2^{n(H(X)+\varepsilon)}$	$\frac{\lvert A_\varepsilon(X)\rvert}{2^n}$
100	$1.17 \cdot 10^{26}$	$7.51 \cdot 10^{27}$	$1.05 \cdot 10^{29}$	$5.9 \cdot 10^{-3}$
500	$1.90 \cdot 10^{131}$	$9.10 \cdot 10^{142}$	$1.34 \cdot 10^{145}$	$2.78 \cdot 10^{-8}$
1000	$4.16 \cdot 10^{262}$	$1.00 \cdot 10^{287}$	$1.79 \cdot 10^{290}$	$9.38 \cdot 10^{-15}$

Example

Example (cont'd)

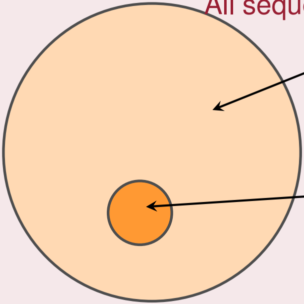
Let $\varepsilon = 0.046$ (5% of $h(1/3)$).

n	$P(\boldsymbol{x} = 11\ldots1)$	$P(\boldsymbol{x} \in A_\varepsilon(X))$
100	$2.4597 \cdot 10^{-18}$	0.660
500	$9.0027 \cdot 10^{-89}$	0.971
1000	$8.1048 \cdot 10^{-177}$	0.998

Source coding

A simple algorithm

Split the message space in two parts and encode separately



Non-typical sequences

$< k^n$ sequences

$\ell_{\boldsymbol{x}} \leq n \log k + 2$ bits

Typical sequences

$\leq 2^{n(H(X)+\varepsilon)}$ sequences

$\ell_{\boldsymbol{x}} \leq n(H(X) + \varepsilon) + 2$ bits

Average codeword length for vector: $E\big[\ell_{\boldsymbol{x}}\big] \rightarrow n\big(H(X) + \delta\big)$

LUND UNIVERSITY

Source coding theorem

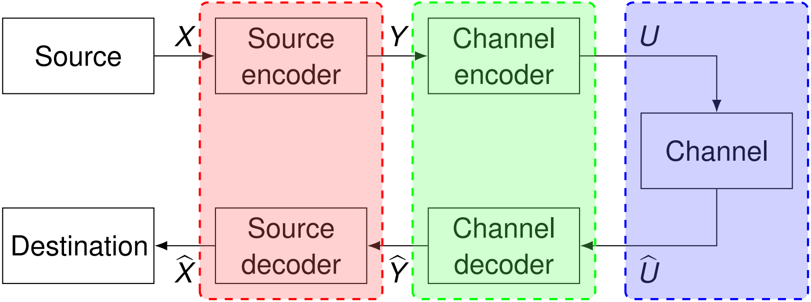
Theorem

Let $\boldsymbol{X} = X_1 \dots X_n$ be a vector of n iid random variables with probability function $p(x)$. Then there exists a code which maps sequences $\boldsymbol{x}$ of length n to binary sequences such that the mapping is invertible and

$$L = \frac{1}{n} E\left[\ell_{\boldsymbol{x}}\right] \leq H(X) + \delta$$

where δ can be made arbitrarily small for sufficiently large n.

Communication system



Channel coding

- Information symbols: $U \in \mathcal{U} = \{u_1, u_2, \dots, u_M\}$
- Encoding function: $x : \mathcal{U} \to \mathcal{X}$. Denote the codewords $x_i = x(u_i), i = 1, \dots M$. For us the codewords are binary vectors of length n, $\mathcal{X} \in \{0, 1\}^n$.
- Channel: Errors occur during transmission and the received symbols are $y \in \mathcal{Y}$. The channel is modeled with $P(Y|X)$.
- Decoding function: $g : \mathcal{Y} \to \mathcal{U}$. Then $\hat{u} = g(y)$. Decoding error if $\hat{u} \neq u$, where u transmitted codeword

The code is an (M, n) code with code rate $R = \frac{\log M}{n} = \frac{k}{n}$.

Discrete memoryless channel (DMC)

Definition

A discrete memoryless channel (DMC) is a system $(\mathcal{X}, P(y|x), \mathcal{Y})$, where

- input alphabet $\mathcal{X}$
- output alphabet $\mathcal{Y}$
- transition probability distribution $P(y|x)$

The channel is memoryless if the probability distribution is independent of previous input symbols.

Channel capacity

Definition

The information channel capacity of a discrete memoryless channel (DMC) is

$$C = \max_{p(x)} I(X;Y)$$

where the maximum is taken over all input distributions.

Theorem

For the DMC $(\mathcal{X}, P(y|x), \mathcal{Y})$*, the channel capacity is bounded by*

$$0 \leq C \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$$

Jointly typical

Definition

The set $A_\varepsilon(X, Y)$ of jointly typical sequences $(\boldsymbol{x}, \boldsymbol{y})$ of length n with respect to the distribution $p(x, y)$ is the set length n sequences

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n) \text{ and } \boldsymbol{y} = (y_1, y_2, \dots, y_n)$$

such that

$$\left| -\frac{1}{n} \log p(\boldsymbol{x}) - H(X) \right| \leq \varepsilon,$$

$$\left| -\frac{1}{n} \log p(\boldsymbol{y}) - H(Y) \right| \leq \varepsilon,$$

$$\left| -\frac{1}{n} \log p(\boldsymbol{x}, \boldsymbol{y}) - H(X, Y) \right| \leq \varepsilon$$

where $p(\boldsymbol{x}, \boldsymbol{y}) = \prod_i p(x_i, y_i)$.

Jointly typical

Equivalent definition

Equivalently, the set $A_\varepsilon(X, Y)$ of jointly typical sequences $(\boldsymbol{x}, \boldsymbol{y})$ can be defined from

$$2^{-n(H(X)+\varepsilon)} \leq p(\boldsymbol{x}) \leq 2^{-n(H(X)-\varepsilon)}$$
$$2^{-n(H(Y)+\varepsilon)} \leq p(\boldsymbol{y}) \leq 2^{-n(H(Y)-\varepsilon)}$$
$$2^{-n(H(X,Y)+\varepsilon)} \leq p(\boldsymbol{x}, \boldsymbol{y}) \leq 2^{-n(H(X,Y)-\varepsilon)}$$

Jointly typical—Properties

Theorem

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be sequences of length n drawn iid according to $p(\boldsymbol{x}, \boldsymbol{y}) = \prod_i p(x_i, y_i)$. Then, for sufficiently large n,

1. $P\Big((\boldsymbol{x}, \boldsymbol{y}) \in A_\varepsilon(X, Y)\Big) \geq 1 - \varepsilon$
2. $(1-\varepsilon)2^{n(H(X,Y)-\varepsilon)} \leq |A_\varepsilon(X, Y))| \leq 2^{n(H(X,Y)+\varepsilon)}$
3. *If $(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}})$ drawn from $p(\boldsymbol{x})p(\boldsymbol{y})$, i.e. $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ are independent with the same marginals as $p(\boldsymbol{x}, \boldsymbol{y})$. Then*

$$(1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)} \leq P\Big((\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \in A_\varepsilon(X, Y)\Big) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

Channel coding theorem

Achievable code rate

A code rate is acheivable if there exists a $(2^{nR}, n)$ code such that the error probability can be made arbitrarily small, i.e.

$$P_e = P(g(\boldsymbol{Y}) \neq u_i | \boldsymbol{X} = x(u_i)) \rightarrow 0, n \rightarrow \infty$$

Theorem (Channel Coding Theorem)

A code rate is achievable if and only if

$$R < C = \max_{p(x)} I(X; Y)$$

Channel coding theorem

Meaning of Channel coding theorem

Consider a dicrete memoryless channel with information capacity C and code prestanda $(2^{nR}, n)$. Then

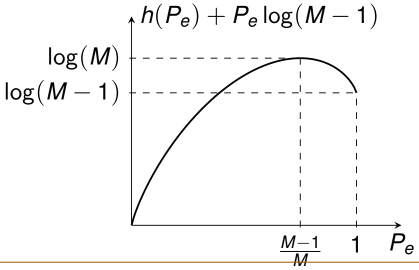
- If $R < C$ it is possible to transmitt information with arbitrarily low error probability.
- If $R > C$ it is not possible to achieve reliable communication.

Fano's lemma

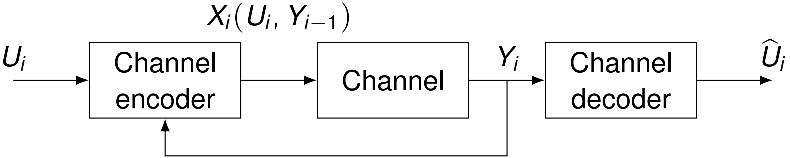
Theorem

If U and $\widehat{U}$ are two stochastic variables over the same alphabet with M letters, and $P_e = P(U \neq \widehat{U})$ is the error probability, then

$$h(P_e) + P_e \log(M-1) \geq H(U|\widehat{U})$$



Channel with feedback



Definition

In a feedback channel a discrete memoryless channel is used, and the previously received symbol y_{i-1} is available at the encoder, i.e. the code symbol at time i is $x(u, y_{i-1})$.

Channel with feedback

Definition

In a feedback channel a discrete memoryless channel is used, and the previously received symbol Y_{i-1} is available at the encoder, i.e. the code symbol at time i is $x(u, y_{i-1})$.

Theorem

The capacity for a feedback channel is equal to the non-feedback channel,

$$C_{FB} = C = \max_{p(x)} I(X; Y)$$