Lecture 1: Elements of Information Theory

Course instructor: Alexey Frolov

al.frolov@skoltech.ru

Teaching Assistant: Stanislav Kruglik

stanislav.kruglik@skolkovotech.ru

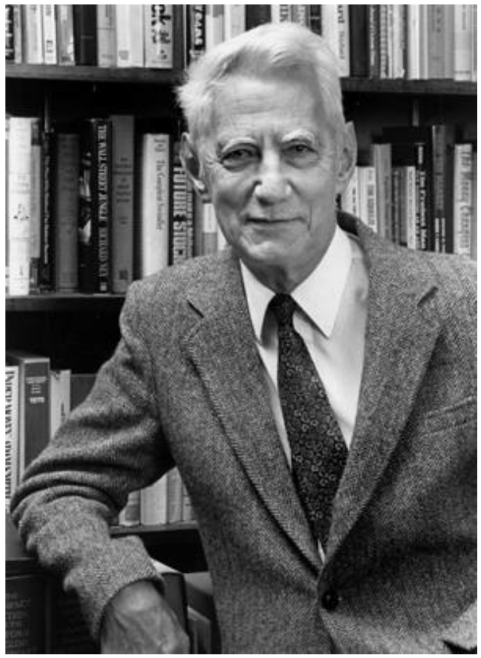
January 31, 2017

Outline

Claude Shannon

Shannon C. E. A mathematical theory of communication. — Bell Syst. Tech. J., 1948, v. 27, p. 379–423 (Part I), p. 623–656 (Part II).

[Рус. пер.: *Шеннон К.* Работы по теории информации и кибернетике. М.: ИЛ, 1963, с. 243–332.]



30.04.1916 — 26.02.2001

Probabilities

Two methods to describe a random variable (R.V.) X

- a function $X : \Omega \to \mathcal{X}$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a target space $\mathcal{X}$
- a distribution P_X on some measurable space $(\mathcal{X}, \mathcal{F})$

Definition (Discrete R.V.)

R.V. X is discrete if if there exists a countable set $\mathcal{X} = \{x_j, j = 1, \ldots\}$, such that

$$\sum_{j=1}^{\infty} P_X(x_j) = 1.$$

In what follows:

- $\mathcal{X}$ – alphabet
- $x \in \mathcal{X}$ – atoms
- P_X – probability mass function (pmf).

Briefer notation will be used. We will write $P(x)$ instead of $P_X(x)$.

By $\mathbb{E}[X]$ we denote the expectation (mean) of X.

Probabilities

Let us consider R.V. X and Y with alphabets $\mathcal{X}$ and $\mathcal{Y}$. By $P(x, y)$ (or $P_{X,Y}(x, y)$) we denote a *joint* probability of x and y.

We can obtain the *marginal* probability $P(x)$ from the joint probability $P(x, y)$ by summation:

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y).$$

Probabilities

Conditional probability:

$$P(x|y) = \frac{P(x,y)}{P(y)}.$$

Independence:

$$X \perp Y \text{ iff } P(x,y) = P(x)P(y).$$

Chain rule

$$P(x,y) = P(x|y)P(y).$$

Sum rule

$$P(x) = \sum_{y \in \mathcal{Y}} P(x|y)P(y).$$

Bayes' theorem

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}.$$

Entropy

How to measure the randomness of R.V. X?

Entropy

Shannon entropy $H(p_1, p_2, \ldots, p_n)$ is characterized by a small number of criteria ($p_i = P(x_i)$).

- **Continuity.** Changing the values of the probabilities by a very small amount should only change the entropy by a small amount.
- **Symmetry.** E.g. $H(p_1, p_2, \ldots, p_n) = H(p_2, p_1, \ldots, p_n)$.
- **Maximum.** The measure should be maximal if all the outcomes are equally likely (uncertainty is highest when all possible events are equiprobable). For equiprobable events the entropy should increase with the number of outcomes.
- **Additivity.** The amount of entropy should be independent of how the process is regarded as being divided into parts.

$$H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = H\left(\frac{b_1}{n}, \ldots, \frac{b_k}{n}\right) + \sum_{i=1}^{k} \frac{b_i}{n} H\left(\frac{1}{b_1}, \ldots, \frac{1}{b_k}\right)$$

Entropy

Any definition of entropy satisfying these assumptions has the form

$$H(X) = -c \sum_{x \in \mathcal{X}} P(x) \log P(x) = c\mathbb{E}[\log(1/P(X))],$$

where c is a constant corresponding to a choice of measurement units. We agree that $0 \log 0 = 0$.

$\log_2 \leftrightarrow$ bits
$\ln \leftrightarrow$ nats
$\log_{256} \leftrightarrow$ bytes
$\log \leftrightarrow$ arbitrary units, base always matches exp

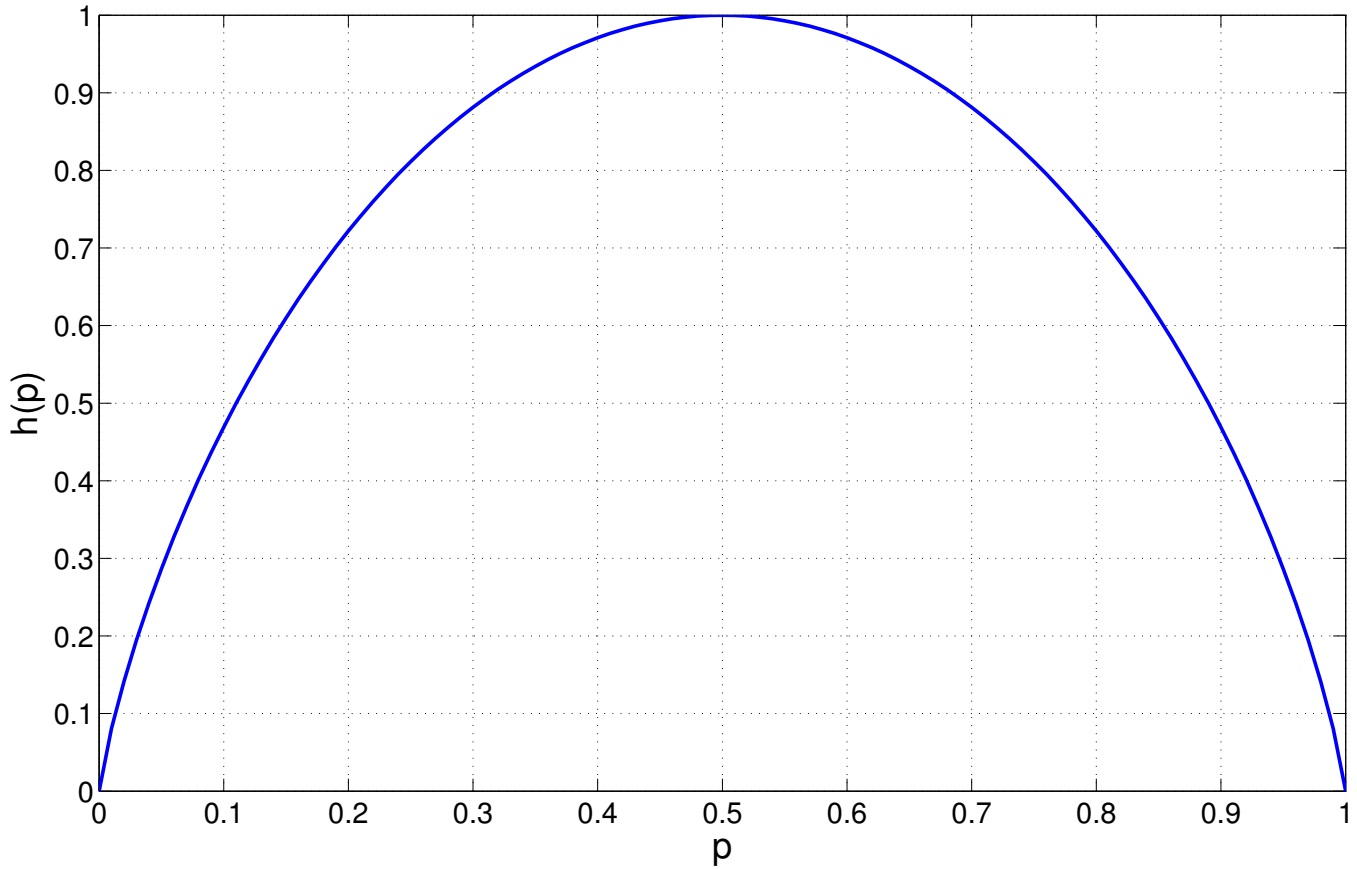
$$H(X) \geq 0$$

Entropy

Let $X \sim \text{Bern}(p)$. Then

$$H(X) = h(p) \triangleq -p \log p - (1-p) \log(1-p),$$

where $h(p)$ – entropy function.



Entropy

Example

$X : \mathcal{X} = \{a, b, c, d\}, P_X = \{1/2, 1/4, 1/8, 1/8\}.$

$$\begin{aligned} H(X) &= -1/2 \log(1/2) - 1/4 \log(1/4) - 2/8 \log(1/8) \\ &= 1.75 \text{ bits} \end{aligned}$$

Suppose we want to determine the value of X with the minimum number of binary questions. The expected number of binary questions is 1.75. The minimal number of questions is in between $H(X)$ and $H(X) + 1$.

Joint entropy

Definition

The *joint* entropy $H(X, Y)$ can be defined as

$$H(X, Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x, y) \log P(x, y) = -\mathbb{E}[\log P(x, y)].$$

Conditional entropy

Definition

The *conditional* entropy $H(Y|X)$ can be defined as

$$\begin{aligned} H(X,Y) &= \sum_{x\in\mathcal{X}} P(x)H(Y|x) \\ &= -\sum_{x\in\mathcal{X}} P(x)\sum_{y\in\mathcal{Y}} P(y|x)\log P(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(y|x) \\ &= \mathbb{E}[\log P(Y|X)]. \end{aligned}$$

Chain rule

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Relative entropy

Entropy is a measure of uncertainty of R.V. Relative entropy is a measure of distance in between two distributions $P(x)$ and $Q(x)$.

Definition

The relative entropy or Kullbak–Leibler distance in between p.m.f.'s $P(x)$ and $Q(x)$ is defined as

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}\left[\frac{P(X)}{Q(X)}\right].$$

We use convention $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

Is it a metric?

Relative entropy

Example

$X \sim \text{Bern}(r)$, $Y \sim \text{Bern}(s)$

$$D(P||Q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$

and

$$D(Q||P) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}$$

If $r = 0.5$ and $s = 0.25$, then $D(P||Q) = 0.2075$ bit and $D(Q||P) = 0.1887$ bit

Check triangle inequality.

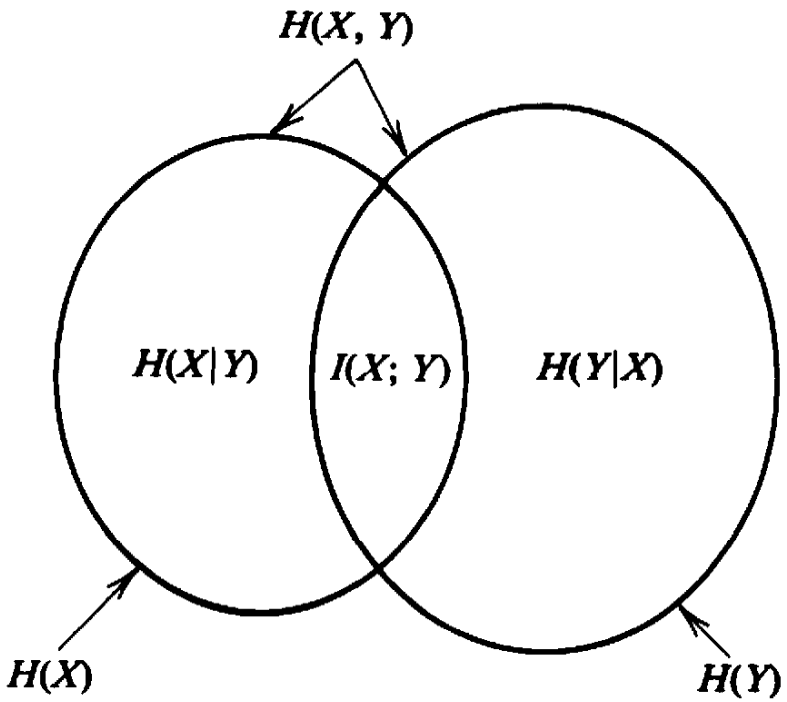
Mutual information

Definition

$$
\begin{aligned}
I(X;Y) &= \sum_{x \in cX} \sum_{y \in cY} P(x,y) \log \frac{P(X,Y)}{P(X)P(Y)} \\
&= D(P_{X,Y} || P_X P_Y) \\
&= \mathbb{E}\left[\log \frac{P(X,Y)}{P(X)P(Y)} \right]
\end{aligned}
$$

Mutual information and entropy

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$



Chain rules

Entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$$

Mutual information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} H(X_i; Y|X_1, \ldots, X_{i-1})$$

Conditional mutual information

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z).$$

Jensen's inequality

Definition

A function $f(x)$ is said to be convex over an interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $\lambda \in [0, 1]$

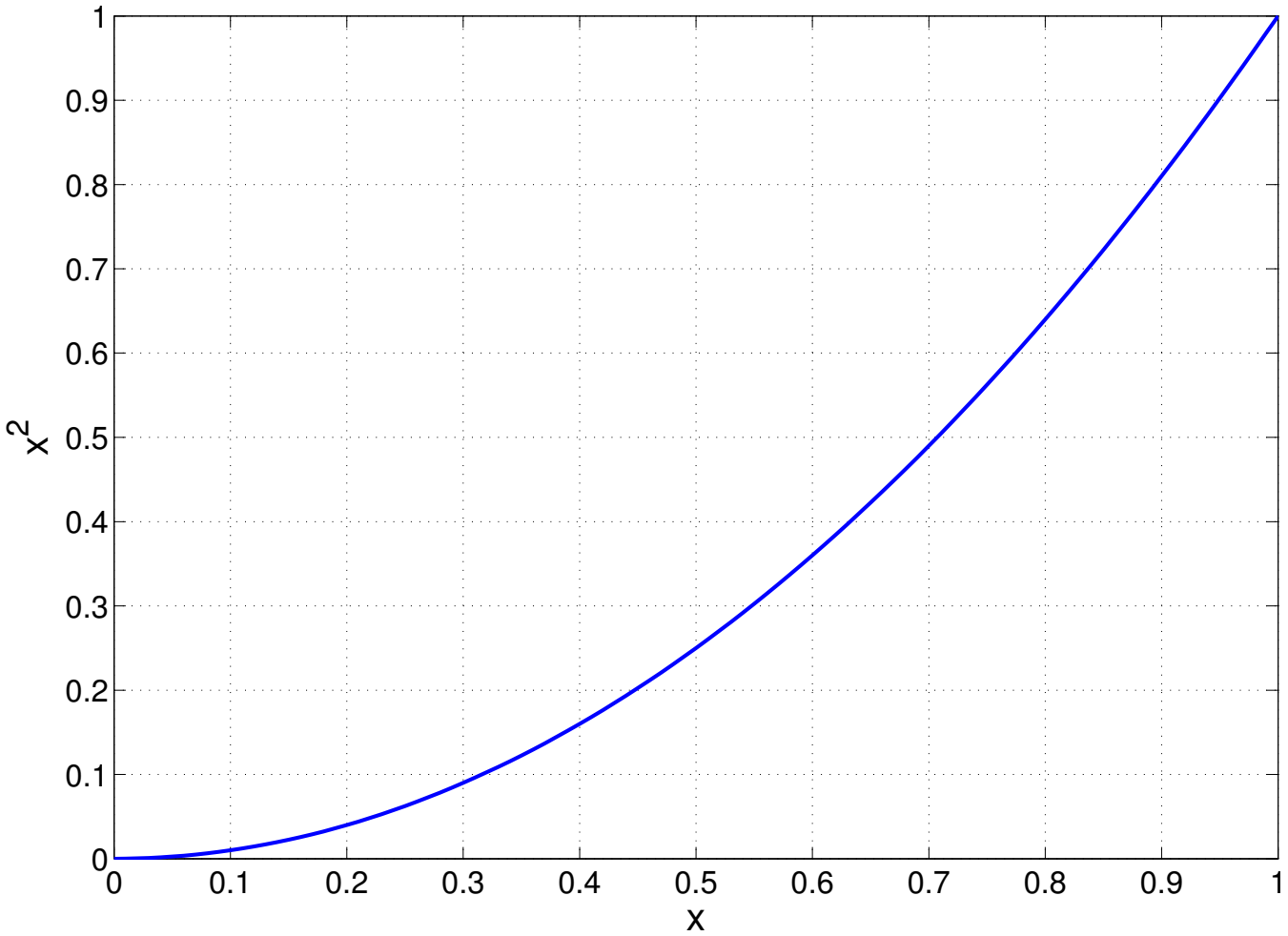
$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Strictly convex is the equality holds only for $\lambda = 0$ or $\lambda = 1$.

Definition

$f(x)$ is concave if $-f(x)$ is convex.

Jensen's inequality



How to check the convexity of a function?

Jensen's inequality

Theorem

If $f(x)$ is a convex function and X is R.V.

$$\mathbb{E}[f(X)] \geq f\left(\mathbb{E}[X]\right).$$

Jensen's inequality

Theorem (Information inequality)

$$D(P||Q) \geq 0$$

Proof.

$$\begin{aligned} -D(P||Q) &= -\sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \\ &= \sum_{x\in\mathcal{X}} P(x)\log\frac{Q(x)}{P(x)} \\ &\leq \log\sum_{x\in\mathcal{X}} P(x)\frac{Q(x)}{P(x)} = 0. \end{aligned}$$

Jensen's inequality

Corollary

$$I(X;Y) \geq 0,$$

$$H(X) \leq \log |\mathcal{X}|,$$

$$H(X|Y) \leq H(X),$$

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i).$$

Log sum inequality

Theorem (Log sum inequality)

For non-negative numbers $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\left(\sum_{i=1}^{n} a_i \right)}{\left(\sum_{i=1}^{n} b_i \right)}$$

Log sum inequality

Corollary

- *$D(P||Q)$ is convex in the pair (P, Q), i.e.*

$$\begin{aligned} D(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2) &\leq \lambda D(P_1||Q_1) \\ &+ (1-\lambda)D(P_2||Q_2). \end{aligned}$$

- *$H(P)$ is a concave function of P.*
- *$I(X;Y)$ is a concave function of $P(X)$ for fixed $P(Y|X)$ and a convex function of $P(Y|X)$ for fixed $P(X)$.*

Data processing inequality

Data processing inequality can be used to show, that no clever manipulation of the data can improve the inferences that can be made from the data.

Definition

R.V.'s X, Y and Z form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and conditionally independent of X, i.e.

$$P(z|y,x) = P(z|y).$$

Consequences:

- $P(x,z|y) = P(x|y)P(z|y)$
- $X \to Y \to Z$ implies $Z \to Y \to X$
- $Z = f(Y)$, then $X \to Y \to Z$

Data processing inequality

Theorem (Data processing inequality)

If $X \to Y \to Z$, *then*

$$I(X;Y) \geq I(X;Z).$$

Proof.

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y).$$

□

$Z = f(Y)$, $I(X;Y) \geq I(X;f(Y))$.

Fano's inequality

Assume X and Y are correlated and we want to guess X given Y. Let $\hat{X} = g(Y)$. We have $X \to Y \to \hat{X}$. By P_e we denote the probability of error

$$P_e = \Pr(X \neq \hat{X}).$$

Theorem (Fano's inequality)

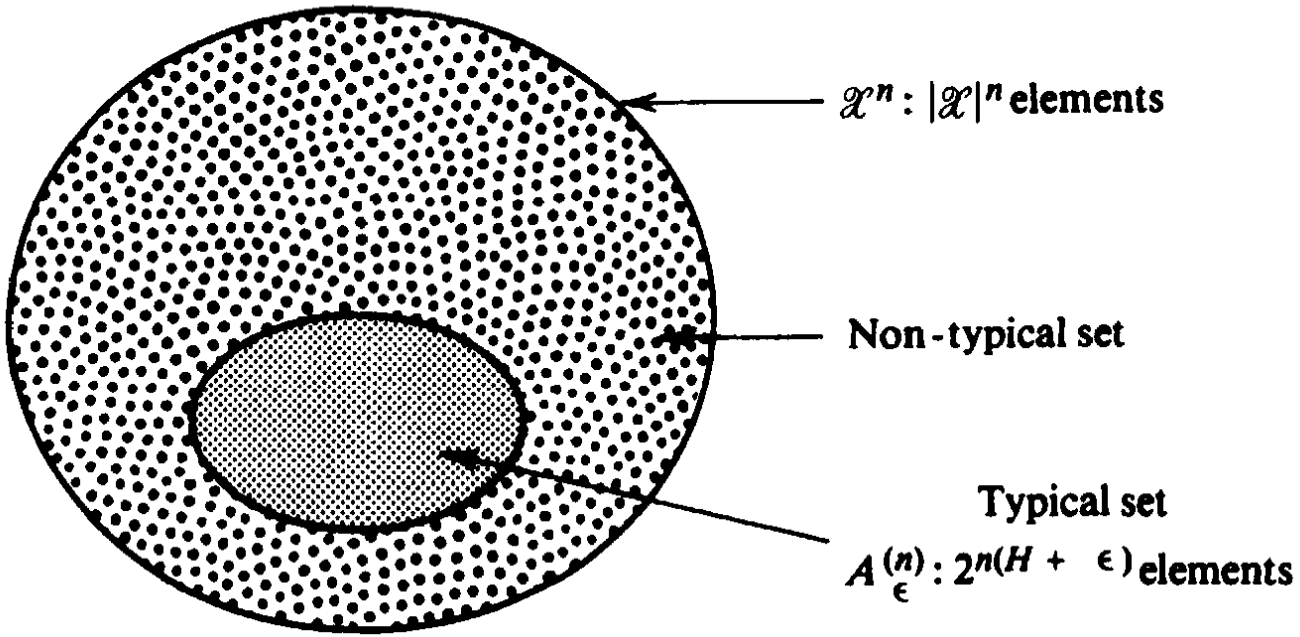
$$h(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

Proof.

By E we denote the indicator of error.

$$H(E, X|Y) = H(X|Y) + H(E|X, Y) = H(E|Y) + H(X|E, Y).$$

Typical set


$\mathcal{X}^n : |\mathcal{X}|^n$ elements
Non-typical set
Typical set
$A_\epsilon^{(n)} : 2^{n(H + \epsilon)}$ elements

Thank you for your attention!