Information Theory. 1st Chapter Slides

Boris Kudryashov

ITMO University

December 18, 2016

Agenda

1. Discrete Sources
2. Information measurement. Self Information.
3. Entropy
4. Convex functions of multiple variables.
5. Conditional Entropy
6. Discrete random sequences. Markov Chains.
7. Entropy per message of discrete stationary source.
8. Uniform coding of discrete source.
9. Chebyshev inequality. The law of large numbers.
10. Forward coding theorem for discrete memoryless source.
11. Reverse coding theorem for discrete memoryless source.
12. Set of typical sequences for discrete DMS. Discrete sources with memory.

- Probability is *of a compound event* A:

$$P(A) = \sum_{x \in A} p(x).$$

- Over Ω, Boolean algebra is defined:

$$\begin{aligned}
&P(\varnothing) = 0;\\
&P(X) = 1;\\
&P(A^c) = 1 - P(A);\\
&P(A \cup B) = P(A) + P(B) - P(AB).
\end{aligned}$$

- Additive assessment of probability of events sum:

$$P(\bigcup_{m=1}^{M} A_m) \le \sum_{m=1}^{M} P(A_m)$$

- Conditional probability:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- For arbitrary number of events:

$$P(A_1...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)...P(A_n|A_1...A_{n-1}).$$

- $A, B \subseteq X$ are independent, if:

$$P(AB) = P(A)P(B).$$

- $A_1, ..., A_n \subseteq X$ are mutually independent, if:

$$P(A_1...A_n) = P(A_1)P(A_2)...P(A_n).$$

- Ff $A, B \subseteq X$ are independent

$$P(A|B) = P(A); P(B|A) = P(B).$$

- Formula of total probability

$$P(A) = \sum_{m=1}^{M} P(A|H_m)P(H_m)$$

- Bayes' law

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{m=1}^{M} P(A|H_m)P(H_m)}$$

- Multiplication of X and Y is

$$Z = XY = \{(x, y) : x \in X, y \in Y\}$$

- Multiplication of ensembles $X = \{x, p_X(x)\}$ and $Y = \{y, p_Y(y)\}$, requires a joint probability distribution $\{p_{XY}(x, y)\}$ on XY. As a result we get $XY = \{(x, y), p_{XY}(x, y)\}$.
- Conditional probability distribution

$$p(x|y) = \begin{cases} \frac{p(x,y)}{p(y)}, & \text{if } p(y) \neq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad x \in X.$$

- Ensembles X and Y are independent, if

$$p(x, y) = p(x)p(y), \quad x \in X, \quad y \in Y.$$

-

$$p(x_1, ..., x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1x_2)...p(x_n|x_1, ..., x_{n-1}).$$

- Mathematical expectation of X:

$$\mathbf{M}_X[x] = \sum_{x \in X} xp(x)$$

- Dispersion

$$\mathbf{D}_X[x] = \mathbf{M}_X\left[(x - \mathbf{M}_X[x])^2\right]$$

- Correlation

$$K_{XY}(x, y) = \mathbf{M}_{XY}\left[(x - \mathbf{M}[x])(y - \mathbf{M}[y])\right].$$

- Mathematical expectation property:

$$\mathbf{M}_Y[y] = \mathbf{M}_X[\varphi(x)] = \sum_{x \in X} \varphi(x) p_X(x). \qquad (1)$$

- Proof:

$$\begin{aligned} \mathbf{M}_Y[y] &= \sum_{y \in Y} y p_Y(y) & = \\ \sum_{y \in Y} y \sum_{x:\varphi(x)=y} p_X(x) &= \sum_{y \in Y} \sum_{x:\varphi(x)=y} y p_X(x) & = \\ \sum_{y \in Y} \sum_{x:\varphi(x)=y} \varphi(x) p_X(x) &= \sum_{x \in X} \varphi(x) p_X(x). \end{aligned}$$

Properties of random variables

- $\mathbf{M}[x+y] = \mathbf{M}[x] + \mathbf{M}[y]$.
- $\mathbf{M}[cx] = c\mathbf{M}[x]$.
- $\mathbf{M}[xy] = \mathbf{M}[x]\mathbf{M}[y]$.
- $\mathbf{D}[x+y] = \mathbf{D}[x] + \mathbf{D}[y]$.
- $\mathbf{D}[cx] = c^2\mathbf{D}[x]$.
- $\mathbf{D}[c+x] = \mathbf{D}[x]$.
- If x and y are independent, then $K(x,y) = 0$. That is, independent random variables are uncorrelated (but not vice versa).

- Requirement to requirement for information measure:

$$\mu(x_1, \dots, x_n) = \mu(x_1) + \dots + \mu(x_n)$$

- Self information $I(x)$ of message x, from $X = \{x, p(x)\}$,

$$I(x) = -\log p(x). \tag{2}$$

Properties of self information

- *Non-negative*: $I(x) \geq 0, x \in X$.
- *Monotone*: if $x_1, x_2 \in X$, $p(x_1) \geq p(x_2)$, то $I(x_1) \leq I(x_2)$
- *Additive*. For independent messages $x_1, ..., x_n$ holds

$$I(x_1, ..., x_n) = \sum_{i=1}^{n} I(x_i).$$

Entropy and It's properties

- Entropy of discrete ensemble $X = \{x, p(x)\}$ is

$$H(X) = \mathbf{M}\left[-\log p(x)\right] = -\sum_{x \in X} p(x) \log p(x) \quad .$$

1 $H(X) \geq 0$.

2 $H(X) \leq \log |X|$. Equality is reached iff elements of X have equal probability.

3 If probability distributions for ensembles X and Y are equal sets of numbers, then holds $H(X) = H(Y)$.

4 Is X and Y are independent,

$$H(XY) = H(X) + H(Y).$$

Entropy and It's properties

5 entropy is convex $\cap$ function of probability distribution on elements of X.

6 Let $X = \{x, p(x)\}$ and $A \subseteq X$. Consider $X' = \{x, p'(x)\}$. Let $p'(x)$ be:

$$p'(x) = \begin{cases} \frac{P(A)}{|A|}, x \in A, \\ p(x), x \notin A. \end{cases}$$

Then $H(X') \geq H(X)$.

7 Consider ensemble X. Let $g(x)$. be defined on X. Consider $Y = \{y = g(x)\}$. Then $H(Y) \leq H(X)$. Equality is achieved when function $g(x)$ is bijective.

Proof of Property (2)

- Consider difference between lhs and rhs:

$$
\begin{aligned}
H(X) - \log|X| &\overset{\text{(a)}}{=} -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X} p(x)\log|X| = \\
&\overset{\text{(b)}}{=} \sum_{x\in X} p(x)\log\frac{1}{p(x)|X|} \le \\
&\overset{\text{(c)}}{\le} \log e\left[\sum_{x\in X} p(x)\left(\frac{1}{p(x)|X|} - 1\right)\right] = \\
&= \log e\left(\sum_{x\in X}\frac{1}{|X|} - \sum_{x\in X} p(x)\right) = 0 \quad .
\end{aligned}
$$

Proof of Property (2)

- $\ln x \leq x - 1 <=> \log x \leq (x-1)\log e$.

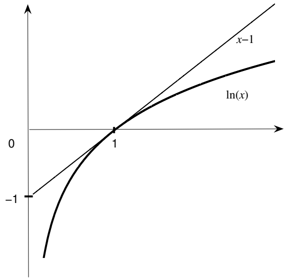

Figure: Graphical interpretation of $\ln(x) \leq x - 1$

Example

- $X = \{0, 1\}$. Let $p(1) = p$, $p(0) = 1 - p = q$.
- Entropy of binary ensemble

$$H(X) = -p \log p - q \log q \eta(p). \qquad (3)$$

- First derivative of $\eta(p)$.

$$\eta'(p) = -\log p + \log(1 - p).$$

- Second derivative of $\eta(p)$.

$$\eta''(p) = -\log e/p - \log e/(1 - p) < 0,$$

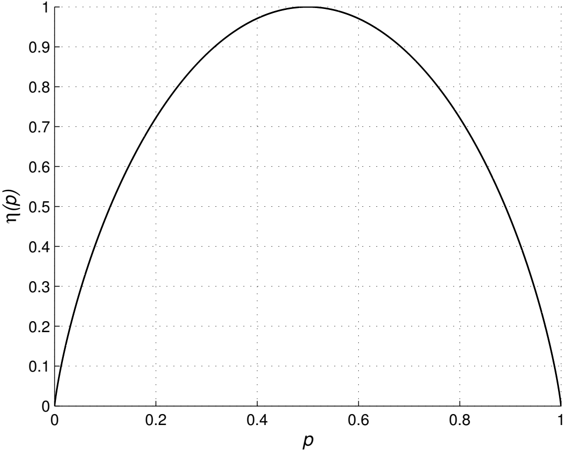

Figure: Entropy of binary ensemble

- Set of real vectors R is convex, if $\forall \boldsymbol{x}, \boldsymbol{x}' \in R$ and $\forall \alpha \in [0,1]$, vector $\boldsymbol{y} = \alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}'$ is in R.

- Theorem

Set of probability vectors of length M is convex.

Proof: $X = \{1, 2, \ldots, M\}$
For $\boldsymbol{p} = (p_1, ..., p_M)$, $\boldsymbol{p}' = (p_1', ..., p_M')$ and $\alpha \in [0,1]$ consider

$$\boldsymbol{q} = \alpha \boldsymbol{p} + (1-\alpha)\boldsymbol{p}'.$$

Sum of $\boldsymbol{q}$ components is

$$\sum_{i=1}^{M} q_i = \alpha \sum_{i=1}^{M} p_i + (1-\alpha) \sum_{i=1}^{M} p_i' = \alpha + 1 - \alpha = 1.$$

- $f(\mathbf{x})$ is convex if $\forall \mathbf{x}, \mathbf{x}' \in R$ and $\forall \alpha \in [0,1]$ holds:

$$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{x}') \geq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{x}') \qquad (4)$$

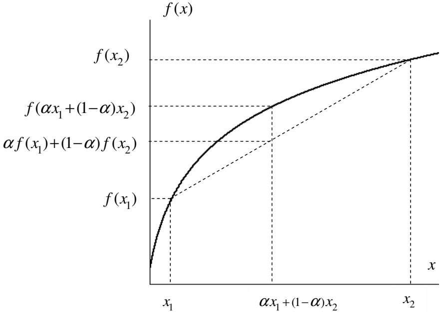

Figure: Определение выпуклой функции

- $$f(\alpha x_1 + (1-\alpha)x_2) \geq \alpha f(x_1) + (1-\alpha) f(x_2),$$

Theorem

Lst $f(\boldsymbol{x})$ be convex $\cap$ function of $\boldsymbol{x}$, defined on a convex set R and let $\alpha_1, ..., \alpha_M \in [0,1]$ be such that $\sum\limits_{m=1}^{M} \alpha_m = 1$. Then $\forall \boldsymbol{x}_1, ..., \boldsymbol{x}_M \in R$ holds

$$f\left(\sum_{m=1}^{M} \alpha_m \boldsymbol{x}_m\right) \geq \sum_{m=1}^{M} \alpha_m f(\boldsymbol{x}_m). \tag{5}$$

Properties of convex functions

1. Sum of convex functions is convex
2. Product of convex function and positive constant is convex function.
3. A linear combination of convex functions with non-negative coefficients is a convex function.

Theorem

Entropy $H(\boldsymbol{p})$ of ensemble with probability distribution $\boldsymbol{p}$ is a convex $\cap$ function of $\boldsymbol{p}$.

Proof. By Entropy definition:

$$H(\boldsymbol{p}) = -\sum_{m=1}^{M} p_m \log p_m = \sum_{m=1}^{M} f_m(\boldsymbol{p}). \qquad (6)$$

Consider $f_m(\boldsymbol{p})$. $f_m''(\boldsymbol{p}) = -(\log e)/p_m$.
$f_m''(\boldsymbol{p}) \leq 0 \forall p_m \in (0,1)$.

Proof of Entropy property (6)

- Denote $\tilde{\boldsymbol{p}} = ((p_1+p_2)/2, (p_1+p_2)/2, p_3, ..., p_M)$.
- We should prove
$$H(\tilde{\boldsymbol{p}}) \geq H(\boldsymbol{p}). \tag{7}$$
- Denote
$$\begin{aligned} \boldsymbol{p}' = \boldsymbol{p} &= (p_1, p_2, p_3, ..., p_M), \\ \boldsymbol{p}'' &= (p_2, p_1, p_3, ..., p_M). \end{aligned}$$
- Note, that: $H(\boldsymbol{p}') = H(\boldsymbol{p}'') = H(\boldsymbol{p})$.
- Holds: $\tilde{\boldsymbol{p}} = (\boldsymbol{p}' + \boldsymbol{p}'')/2$.
- From entropy convexity:
$$\begin{aligned} H(\tilde{\boldsymbol{p}}) &= H\left(\frac{\boldsymbol{p}' + \boldsymbol{p}''}{2}\right) \geq \\ &\geq \frac{1}{2}H(\boldsymbol{p}') + \frac{1}{2}H(\boldsymbol{p}'') = H(\boldsymbol{p}). \end{aligned}$$

- Conditional self information of x when y is fixed:

$$I(x|y) = -\log p(x|y),$$

- Conditional entropy of X when $y \in Y$ is fixed:

$$H(X|y) = -\sum_{x \in X} p(x|y) \log p(x|y), \qquad (8)$$

- Conditional entropy of X when Y is fixed:

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y)$$

Properties of conditional entropy:

1.
$$H(X|Y) \geq 0.$$

2.
$$H(X|Y) \leq H(X),$$

Equality is reached iff X and Y are independent.

3.
$$H(XY) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Properties of Conditional Entropy:

4

$$\begin{aligned} H(X_1 ... X_n) = H(X_1) &+ H(X_2|X_1) + \\ &+ H(X_3|X_1X_2) + ... + \\ &+ H(X_n|X_1, ..., X_{n-1}). \end{aligned}$$

5

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

Equality is achieved iff X and Z are conditionally independent $\forall y \in Y$.

6

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

Equality is achieved iff $X_1, \ldots, X_n$ are mutually independent.

Proof of property (2)

- $$\begin{aligned}
H(X|Y) - H(X) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x|y) + \\
&\quad + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) = \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x)}{p(x|y)} \le \\
&\le \sum_{x\in X}\sum_{y\in Y} p(x,y)\left(\frac{p(x)}{p(x|y)} - 1\right)\log e = \\
&= \left(\sum_{x\in X}\sum_{y\in Y} p(y)p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\right)\log e = \\
&= 0.
\end{aligned}$$

Proof of property (2)

- $$H(X|Y) \overset{(a)}{=} \mathbf{M}_Y\left[H(\boldsymbol{p}_{X|y})\right] \le \overset{(b)}{\le} H\left(\mathbf{M}_Y\left[\boldsymbol{p}_{X|y}\right]\right) = \overset{(c)}{=} H\left(\boldsymbol{p}_X\right) = H(X),$$

- $$\mathbf{M}_Y\left[p(x|y)\right] = \sum_y p(x|y)p(y) = p(x).$$

Proof of property (3) and (4)

- $$p(x, y) = p(x)p(y|x) = p(y)p(x|y),$$

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

Proof of property (5)

- Consider $XYZ = \{(x, y, z), p(x, y, z)\}$. Let $p(x, z|y)$ and $p(x|y)$ be defined.
- $$H(X|y, Z) = \mathbf{M}_{XZ|y}[-\log p(x|yz)],$$
- $$H(X|y) = \mathbf{M}_{X|y}[-\log p(x|y)].$$
- by property (2)

$$H(X|y, Z) \leq H(X|y).$$

Proof of Entropy property (7)

- Consider $X = \{x, p(x)\}$, $g(x)$, $Y = \{y = g(x), x \in X\}$.
- Prove, that
$$H(Y) \leq H(X). \tag{9}$$
- By entropy property
$$H(XY) = \underbrace{H(X|Y)}_{\geq 0} + H(Y) = \underbrace{H(Y|X)}_{=0} + H(X). \tag{10}$$
- As long as $g(x)$ is defined on each x, We have $H(Y|X) = 0$. $H(X|Y) \geq 0$.

- If elements of random sequence are real values, such sequence is called stochastic processes.
- Assume, that values of stochastic process are independent and equally distributed at any moment. Then holds:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i),$$

Where $p(x_i)$ is a probability for $x_i \in X$ to appear at moment i.

- Process is Stationary, if $\forall n, t$ holds

$$p(x_1, ..., x_n) = p(x_{1+t}, ..., x_{n+t}),$$

  where $x_i = x_{i+t}$, $i = 1, \ldots, n$.
- Discrete source, which generates such a stationary process is called Discrete Memoryless Source (DMS).

- Random process $x_1, x_2, \ldots$ is called Markov Chain of connectivity s, if $\forall n$ and $\forall \mathbf{x} = (x_1, ..., x_n) \in X^n$ holds

$$p(\mathbf{x}) = p(x_1, ..., x_s)p(x_{s+1}|x_1, ..., x_s)p(x_{s+2}|x_2 \ldots x_{s+1}$$

$$\times p(x_n|x_{n-s}, \ldots, x_{n-1}).$$

- Markov Process of connectivity s is a random process such that $\forall n > s$ holds:

$$p(x_n|x_1, ..., x_{n-1}) = p(x_n|x_{n-s}, ..., x_{n-1}),$$

- Markov process is defined by initial probability distribution on sequences of first s values (states) and by conditional probabilities $p(x_n|x_{n-s},...,x_{n-1})$ for arbitrary sequences $(x_{n-s},...,x_n)$.
- If conditional probabilities are unchanged after sequence shifts $(x_{n-s},...,x_n)$ by time, such Markov Chain is called Homogeneous.
- Simple Markov Chain is a Homogeneous Markov Chain with $s = 1$ connectivity.
- For Simple Markov Chain definition, states $X = \{0, 1, ..., M-1\}$, initial probability distribution $\{p(x_1), x_1 \in X\}$ and transition probabilities

$$\pi_{ij} = P(x_t = j | x_{t-1} = i), \quad i, j = 0, ..., M-1,$$

are required.

- $M \times M$ probability transition matrix for a Markov Chain

$$\Pi = \begin{bmatrix} \pi_{00} & \pi_{01} & \cdots & \pi_{0,M-1} \\ \pi_{10} & \pi_{11} & \cdots & \pi_{1,M-1} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{M-1,0} & \pi_{M-1,1} & \cdots & \pi_{M-1,M-1} \end{bmatrix}$$

- Consider stochastic vector $\boldsymbol{p}_t = (p_t(0), ..., p_t(M-1))$, which represents Markov Chain states at moment t.
- $p_{t+1}(i) = \sum_{j=1}^{L} p_t(j)\pi_{ji}$
- $\boldsymbol{p}_{t+1} = \boldsymbol{p}_t \Pi$
- For arbitrary number of steps:

$$\boldsymbol{p}_{t+n} = \boldsymbol{p}_t \Pi^n.$$

- Assume, that $\exists \boldsymbol{p}$:

$$\boldsymbol{p} = \boldsymbol{p}\Pi. \tag{11}$$

  Such $\boldsymbol{p}$ is called Stationary Distribution for Markov Chain.
- Final probability distribution is called

$$\boldsymbol{p}_\infty = \lim_{t\to\infty} \boldsymbol{p}_\mathbf{t} = \lim_{t\to\infty} \boldsymbol{p}_1 \Pi^t \tag{12}$$

- Consider Discerete stationary source, which generates $(x_1, x_2, \ldots, x_t, \ldots)$,$x_t \in X_t = X$.
- $H(X_t) = H(X)$ is independent of time.
- Entropy per character of sequence of length n

$$H_n(X) = \frac{H(X^n)}{n},$$

- For a conditional entropy:

$$H(X_n|X_1, \ldots, X_{n-1}) = H(X|X^{n-1}).$$

Theorem

For a Discrete Stationary Source holds:

A. $H(X|X^n)$ *does not increase with increasing* $\boldsymbol{n}$;

B. $H_n(X)$ *does not increase with increasing* $\boldsymbol{n}$;

C. $H_n(X) \geq H(X/X^{n-1})$;

D. $\lim\limits_{n\to\infty} H_n(X) = \lim\limits_{n\to\infty} H(X|X^n)$.

Proof of Theorem

- verify the validity of C.

$$H(X^n) = H(X) + H(X|X^1) + ... + H(X|X^{n-1}).$$

- verify the validity of B.

$$\begin{aligned}
H(X^{n+1}) &\overset{\text{(a)}}{=} H(X_1 ... X_n X_{n+1}) = \\
&\overset{\text{(b)}}{=} H(X_1 ... X_n) + H(X_{n+1}|X_1, ..., X_n) = \\
&\overset{\text{(c)}}{=} H(X^n) + H(X|X^n) \leq \\
&\overset{\text{(d)}}{\leq} H(X^n) + H(X|X^{n-1}) \leq \\
&\overset{\text{(e)}}{\leq} H(X^n) + H_n(X) = \\
&\overset{\text{(f)}}{=} (n+1)H_n(X).
\end{aligned}$$

Proof of Theorem

- Veryfy D. From C follows

$$\lim_{n\to\infty} H_n(X) \geq \lim_{n\to\infty} H(X|X^n). \tag{13}$$

- $\forall n, m \in N, m < n$ holds

$$\begin{aligned}
H(X^n) &= H(X_1 \dots X_n) = \\
&\stackrel{\text{(a)}}{=} H(X_1 \dots X_m) + H(X_{m+1} \dots X_n | X_1, \dots, X_m) = \\
&\stackrel{\text{(b)}}{=} mH_m(X) + H(X_{m+1}|X_1, \dots, X_m) + \dots \\
&\quad + H(X_n|X_1, \dots, X_{n-1}) \leq \\
&\stackrel{\text{(c)}}{\leq} mH_m(X) + (n-m)H(X|X^m).
\end{aligned}$$

Proof of Theorem

- $\forall m$ holds

$$\lim_{n\to\infty} H_n(X) \leq H(X|X^m),$$

- Tend $m- > \infty$

$$\lim_{n\to\infty} H_n(X) \leq \lim_{m\to\infty} H(X|X^m). \tag{14}$$

From (13) и (14) we get necessary statement.

- Denote

$$H_\infty(X) = \lim_{n\to\infty} H_n(X),$$

$$\mathrm{H}(\mathrm{X}|X^\infty) = \lim_{n\to\infty} H(X|X^n).$$

- Consider examples from DMS and Markev Source.

Example Discrete Memoryless Source

- $H(X_1 ... X_n) = H(X_1) + ... + H(X_n)$.
- $H(X^n) = nH(X)$.
- $H_n(X) = H(X)$,
- $H_\infty(X) = H(X)$.
- $H(X|X^n) = H(X_{n+1}|X_1, ..., X_n) = H(X)$,
- $H(X|X^\infty) = H(X)$.

Discrete stationary source

Example for Markov Source

- $H(X|X^n) = H(X_{n+1}|X_1, ..., X_n) =$
  $= \mathrm{H}(\mathrm{X}_{n+1}|X_{n-s+1}, ..., X_n) = H(X|X^s)$.
- $H(X|X^\infty) = H(X|X^s)$.
-

$$\begin{aligned} H(X^n) &= H(X_1 \dots X_s X_{s+1} \dots X_n) = \\ &= H(X_1 \dots X_s) + H(X_{s+1} \dots X_n | X_1, \dots, X_s) \end{aligned} \tag{15}$$

-

$$\begin{aligned} H(X_{s+1} \dots X_n | X_1, \dots, X_s) &= H(X_{s+1}|X_1, \dots, X_s) + \\ &+ H(X_{s+2}|X_1, \dots, X_{s+1}) + \dots \\ &+ H(X_n|X_1, \dots, X_{n-1}), \end{aligned}$$

Example for Markov Source

- $H(X_{s+1}...X_n|X_1,...,X_s) = (n-s)H(X|X^s)$.
- $$H(X^n) = sH_s(X) + (n-s)H(X|X^s). \qquad (16)$$
- $H_\infty(X) = H(X|X^s)$.
- $$\begin{aligned} H_n(X) &= H(X|X^s) + \frac{s}{n}(H_s(X) - H(X|X^s)) = \\ &= H(X|X^n) + \frac{s}{n}(H_s(X) - H(X|X^s)). \end{aligned}$$

- messages $x_1, x_2, ..., x_i \in X$, $i = 1, 2, ...$
- Uniform code rate

$$R = \frac{\lceil \log |C| \rceil}{N} \text{ (bit / character)}, \qquad (17)$$

- Consider set of all sequences of length n, i.e. $C = A^n = \{0, 1\}^n$

$$R = \frac{n}{N} \text{ (бит / букву источника)}.$$

- Bijective encoding is only possible iff

$$|X|^N \leq |C| \tag{18}$$

or

$$R \geq \log |X| \geq H(X).$$

- Probability of decoding error:

$$P_e = P\left(\boldsymbol{x} \notin T\right)$$

Table: Uniform code example

Sequence	Probability	Codeword
aa	1/4	000
ab	1/6	001
ac	1/12	010
ba	1/6	011
bb	1/9	100
bc	1/18	101
ca	1/12	110
cb	1/18	111
cc	1/36	111

- Consider $X = \{x, p(x)\}$. Let $\forall x \in X, x > 0$. Let $P(x \geq A)$ for some $A > 0$.
- $$P(x \geq A) = \sum_{x \geq A} p(x) \leq \sum_{x \geq A} \frac{x}{A} p(x) \leq \frac{1}{A} \sum_{x \in X} x p(x) = \frac{\mathbf{M}[x]}{A}.$$
- Denote $m_x = \mathbf{M}[x]$. Rewrite:

$$P(x \geq A) \leq \frac{m_x}{A}. \tag{19}$$

Chebyshev inequality

- Let $X = \{x, p(x)\}$ e arbitrary random variable. For an arbitrary $\varepsilon > 0$ estimate $P\left(|x - m_x| \geq \varepsilon\right)$. Let $y = |x - m_x|$.

$$P\left(y \geq \varepsilon\right) = P\left(y^2 \geq \varepsilon^2\right) \leq \frac{\mathbf{M}\left[y^2\right]}{\varepsilon^2} = \frac{\mathbf{M}\left[(x - m_x)^2\right]}{\varepsilon^2} = \frac{\mathbf{D}[x]}{\varepsilon^2}.$$

- Chebyshev inequality

$$P\left(|x - m_x| \geq \varepsilon\right) \leq \frac{\sigma_x^2}{\varepsilon^2}, \tag{20}$$

where $\sigma_x^2 = \mathbf{D}[x]$.

Chebyshev inequality

- we are interested in:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - m_x\right| \geq \varepsilon\right)$$

- Let $y = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\mathbf{M}[y] = m_x, \quad \mathbf{D}[y] = \frac{1}{n}\sigma_x^2.$$

- Chebyshev inequality for sums of independent random quantities

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - m_x\right| \geq \varepsilon\right) \leq \frac{\sigma_x^2}{n\varepsilon^2} \tag{21}$$

Theorem

Forward coding theorem let H be entropy of discrete memoryless source. $\forall \varepsilon, \delta > 0 \; \exists n_0$ such that $\forall n > n_0$ there exists uniform cod, which encodes the source by blocks of length n and has code rate $R \leq H + \delta$ and error probability $P_e \leq \varepsilon$.

Proof of Forward coding theorem

- Chose $T \subseteq X^n$:

$$T = \left\{ \boldsymbol{x} : \left| \frac{1}{n} I(\boldsymbol{x}) - H \right| \le \delta_0 \right\}, \tag{22}$$

where $I(\boldsymbol{x}) = -\log p(\boldsymbol{x})$ is self information of $\boldsymbol{x} \in X^n$, and $\delta_0 > 0$

- from (22) follows

$$2^{-n(H+\delta_0)} \le p(\boldsymbol{x}) \le 2^{-n(H-\delta_0)}. \tag{23}$$

- Note, that

$$1 \ge P(T) = \sum_{\boldsymbol{x} \in T} p(\boldsymbol{x}) \ge |T| \min_{\boldsymbol{x} \in T} p(\boldsymbol{x}) \ge |T| 2^{-n(H+\delta_0)}.$$

Proof of Forward coding theorem

- Consequently,

$$|T| \le 2^{n(H+\delta_0)}. \tag{24}$$

- Core rate will be

$$R = \frac{\lceil \log |T| \rceil}{n} \le H + \delta_0 + \frac{1}{n}. \tag{25}$$

- For P_e holds:

$$\begin{aligned} P_e = P(\boldsymbol{x} \notin T) &= P\left(\left| \frac{1}{n} I(\boldsymbol{x}) - H \right| > \delta_0 \right) = \\ &= P\left(\left| \frac{1}{n} \sum_{i=1}^{n} I(x_i) - H \right| > \delta_0 \right) \end{aligned} \tag{26}$$

Proof of Forward coding theorem

- Note, that $\mathbf{M}\left[I(x)\right] = H$.
- Apply Chebyshev inequality to (26)

$$P_e \le \frac{\mathbf{D}\left[I(x)\right]}{n\delta_0^2}. \tag{27}$$

- Let $\delta_0 = \delta/2$.
- When $n \ge n_{01} = \mathbf{D}[I(x)]/(\delta^2\varepsilon)$, from (27): $P_e \le \varepsilon$.
- From (25) : $n \ge n_{02} = 2/\delta$, $R < H + \delta$.
- When $n \ge n_0 = \max(n_{01}, n_{02})$, then code rate R and P_e satisfy the theorem requirements.

Theorem

Reverse coding theorem For a Discrete memoryless source with entropy H $\exists\varepsilon > 0$ *such, that* $\forall\delta > 0$ *and for all uniform code with code rate* $R \leq H - \delta$, *probability of error satisfies* $P_e \geq \varepsilon$.

Proof of Reverse coding theorem

- Code rate is $R = \lceil \log |T_1| \rceil / n$, thus:

$$|T_1| \le 2^{nR} \le 2^{n(H-\delta)} \tag{28}$$

- Probability of correct coding

$$P_c = 1 - P_e = \sum_{\boldsymbol{x} \in T_1} p(\boldsymbol{x}). \tag{29}$$

- Consider auxiliary set

$$T = \left\{ \boldsymbol{x} : \left| \frac{1}{n} I(\boldsymbol{x}) - H \right| \le \delta_0 \right\}, \tag{30}$$

where $I(\boldsymbol{x}) = -\log p(\boldsymbol{x})$ is self information of $\boldsymbol{x} \in X^n$, and $\delta_0 > 0$

Proof of Reverse coding theorem

- split sum in (29) to 2 sums

$$P_c = \sum_{\boldsymbol{x} \in T_1 \cap T} p(\boldsymbol{x}) + \sum_{\boldsymbol{x} \in T_1 \cap T^c} p(\boldsymbol{x}), \qquad (31)$$

- estimate the second sum:

$$\sum_{\boldsymbol{x} \in T_1 \cap T^c} p(\boldsymbol{x}) \le \sum_{\boldsymbol{x} \in T^c} p(\boldsymbol{x}) = P(T^c) = P(\boldsymbol{x} \notin T).$$

- Use Chebyshev inequality

$$\sum_{\boldsymbol{x} \in T_1 \cap T^c} p(\boldsymbol{x}) \le \frac{\mathbf{D}\left[I(x)\right]}{n\delta_o^2} \quad . \qquad (32)$$

Proof of Reverse coding theorem

- For first of sums use $|T_1 \cap T| \le |T_1|$:

$$\begin{aligned}
\sum_{\boldsymbol{x} \in T_1 \cap T} p(\boldsymbol{x}) &\overset{\text{(a)}}{\le} |T_1 \cap T| \max_{\boldsymbol{x} \in T_1 \cap T} p(\boldsymbol{x}) \le \\
&\overset{\text{(b)}}{\le} |T_1| \max_{\boldsymbol{x} \in T_1 \cap T} p(\boldsymbol{x}) \le \\
&\overset{\text{(c)}}{\le} |T_1| \max_{\boldsymbol{x} \in T} p(\boldsymbol{x}). \qquad (33)
\end{aligned}$$

- Substitute (28) and (23) to (33)

$$\sum_{\boldsymbol{x} \in T_1 \cap T} p(\boldsymbol{x}) \le 2^{n(H-\delta)} 2^{-n(H-\delta_0)} = 2^{-n(\delta-\delta_0)}. \qquad (34)$$

Proof of Reverse coding theorem

- Substitute (32) and (34) to (31)

$$P_c \leq 2^{-n(\delta-\delta_0)} + \frac{\mathbf{D}\left[I(x)\right]}{n\delta_o^2} \quad . \tag{35}$$

-

$$|T_1| \geq |X|^n.$$

- Code rate should be

$$R \geq \log|T_1|/n \geq \log|X|. \tag{36}$$

- Use $\varepsilon = \min\{\varepsilon_0, 1/2\}$.
- $\forall n = 1, 2, \ldots$ holds $P_e \geq \varepsilon$

- Average Self information

$$T_n(\delta) = \left\{ \mathbf{x} : \left| \frac{1}{n} I(\mathbf{x}) - H(X) \right| \leq \delta \right\}, \qquad (37)$$

- Theorem

$\forall \delta > 0$ *holds:*

1. $$\lim_{n \to \infty} P\left(T_n(\delta)\right) = 1.$$

2. $\forall n \in N$ *holds:*

$$|T_n(\delta)| \leq 2^{n(H(X)+\delta)}.$$

3. $\forall \varepsilon > 0 \; \exists n_0$ *such, that* $\forall n \geq n_0$ *holds*

$$|T_n(\delta)| \geq (1-\varepsilon) 2^{n(H(X)-\delta)}.$$

Proof of theorem

- First statement follows from (26) and (27).
- Second statement is equivalent to (24).
- Fourth statement is equivalent to (23).
- from First follws, that $\forall \varepsilon > 0$ n_0 such, that for $n > n_0$ holds
$$P\left(T_n(\delta)\right) \geq 1 - \varepsilon. \tag{38}$$
- Estimate $T_n(\delta)$ and apply Fourth statement

$$P\left(T_n(\delta)\right) \leq |T_n(\delta)| \max_{\boldsymbol{x} \in T_n(\delta)} p(\boldsymbol{x}) \leq |T_n(\delta)| \, 2^{-n(H(X)-\delta)}. \tag{39}$$

- For DMS probability of $\boldsymbol{x} = (x_1, ..., x_n)$

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i) = \prod_{x \in X} p(x)^{\tau_x(\boldsymbol{x})}.$$

- Self information of per character:

$$\frac{1}{n} I(\boldsymbol{x}) = -\sum_{x \in X} \frac{\tau_x(\boldsymbol{x})}{n} \log p(x).$$

This value is close to entropy $H(X)$, if

$$\frac{\tau_x(\boldsymbol{x})}{n} \approx \; p(x)$$

- $\forall m$ set of uniquely encodable sequences is a set of sequences x, for which holds:

$$\frac{1}{n} I(\boldsymbol{x}) \approx H(X|X^m)$$

or

$$\frac{1}{n}\left(I(x_1,\ldots,x_m) + \sum_{i=m+1}^{n} I(x_i|x_{i-m},\ldots,x_{i-1})\right) \approx H(X|X^m).$$