Information theory

Some basic definitions: The case of discrete random variables

---

Georgios Ropokis

CentaraleSupélec, Campus Rennes

Course material based on textbook: "Elements of Information Theory", Wiley New York, 1991, by T. M. Cover, and J. A. Thomas.

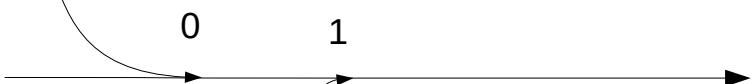
Table of contents

Discrete random variables

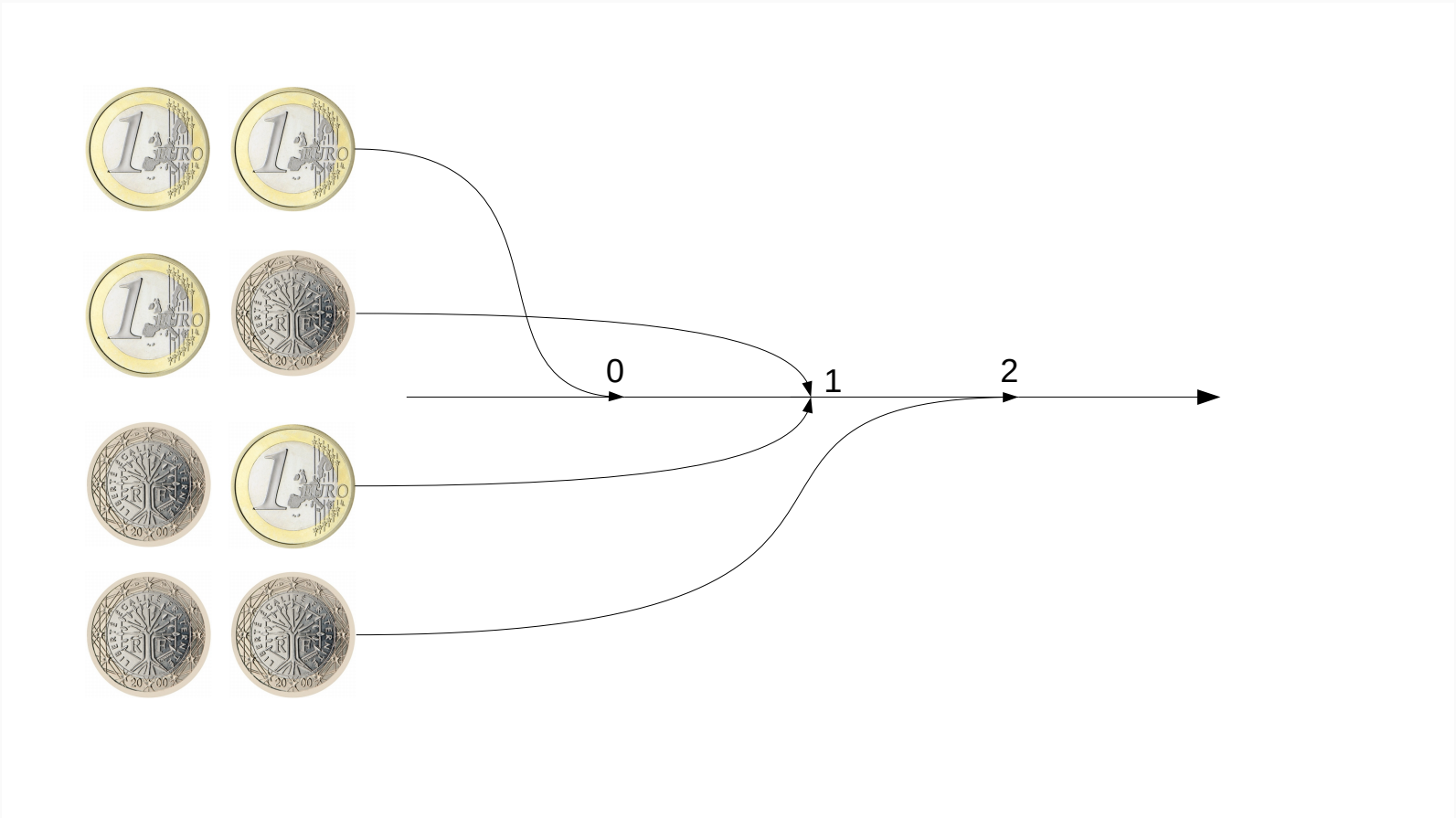
Definition and properties of random variables

- One can define a real random variable as a function, (or a mapping) from the outcome of a random experiment to the set of real numbers.
- In case that the resulting variable takes values on a finite or countably infinite set, we refer to the resulting random variable as a **discrete random variable**.

Example: Toss of a coin



Example: Consecutive tosses of a coin



The mapping does not have to be one-to-one!!

Example: Digital sensor

The outcome of the experiment can be directly a random variable!

Characterizing a discrete random variable

Definition: Probability mass function

Given a discrete random variable X taking values over the set $\mathcal{X} = \{x_1, x_2, \ldots, ...\}$ we define as the probability mass function $p_X[x_i]$, the function that, for each possible value x_i, gives us the probability $\Pr(X = x_i)$.

Definition: Cumulative distribution function

We define the cumulative distribution function $F_X(x)$ as:

$$F_X(x) = \Pr(X \leq x). \tag{1}$$

Remark

Both the probability mass function and the cumulative distribution function uniquely characterize a random variable.

Some well known discrete distributions

Bernoulli distribution

$$p_X[k] = \begin{cases} 1-p, & k=0 \\ p, & k=1. \end{cases} \tag{2}$$

Binomial distribution

$$p_X[k] = \binom{M}{k} p^k (1-p)^{M-k}, k = 0, 1, \ldots, M \tag{3}$$

Geometric distribution

$$p_X[k] = (1-p)^{k-1} p, \quad k = 1, 2, \ldots \tag{4}$$

Poisson distribution

$$p_X[k] = \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots, \quad \text{and} \ \ \lambda > 0 \tag{5}$$

Implicit characterization of random variables

Expectation of a random variable

We define the expectation $\mathbb{E}\{X\}$ of random variable X as:

$$\mathbb{E}\{X\} = \sum_{x_i \in \mathcal{X}} x_i p_X[x_i]. \tag{6}$$

Definition: Expectation of a function of a random variable

For a random variable $Y = g(X)$, we define the expectation of Y as:

$$\mathbb{E}\{Y\} = \mathbb{E}\{g(X)\} = \sum_{x_i \in \mathcal{X}} g(x_i) p_X[x_i]. \tag{7}$$

Definition: Variance of a random variable

$$var(X) = \mathbb{E}\left\{(X - \mathbb{E}\{X\})^2\right\} \tag{8}$$

Entropy

Introduction

- It is defined for any probability mass function, i.e. for any random variable.
- It quantifies the uncertainty of a random variable.
- In that sense, it measures the information carried by this random variable: The highest the uncertainty of a random variable the most the information that we obtain by observing it.

Definition of Entropy

Definition: Entropy

Let X be a discrete random variable taking values over an alphabet $\mathcal{X}$ and having a probability mass function $p_X[x] = \Pr(X = x), x \in \mathcal{X}$. We then define the entropy of random variable X, and denote it as $H(X)$, the quantity:

$$H(X) = -\sum_{x \in \mathcal{X}} p_X[x] \log(p_X[x]) = -\mathbb{E}\{\log(p_X[x])\}. \tag{9}$$

Remarks

- For the logarithmic function involved in the calculation of the entropy, different bases can be used. If the selected base is 2, then we measure entropy in bits, while if the selected base is e, then we measure entropy in nats.
- It is a functional of the probability distribution and does not depend on the values of the random variable itself.

Properties of entropy

Non-negativity property

Since for the values of the probability mass function it holds that $0 \leq p_X[x] \leq 1$, it holds that $\log(p_X[x]) \leq 0$. As a result $H(X)$ is non-negative.

Change of base

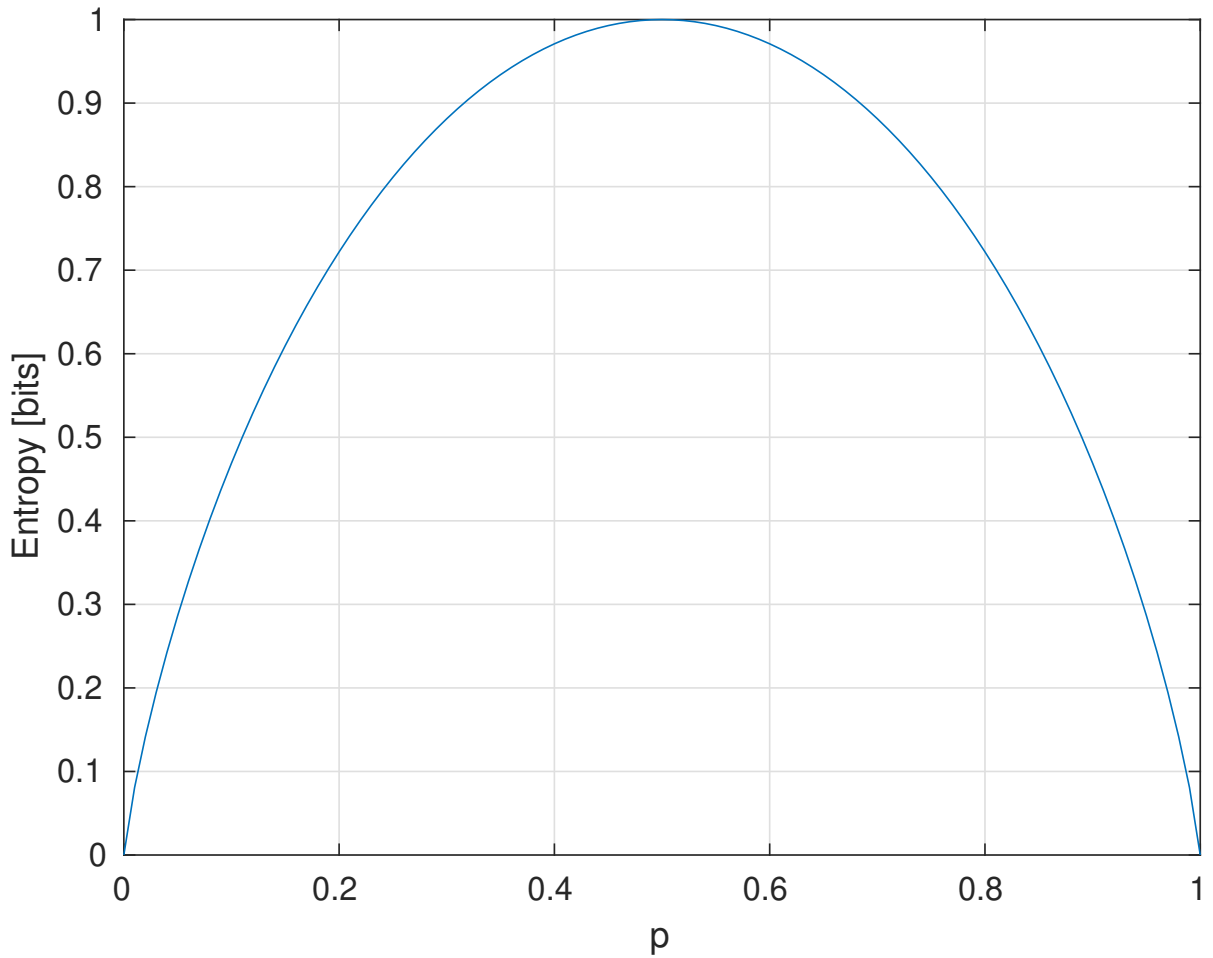
Let $H_b(X) = -\mathbb{E}\{\log_b(p_X[x])\}$. Exploiting the logarithmic property $\log_b z = \log_b a \log_a z$, we obtain that:

$$H_b(X) = \log_b a H_a(X). \tag{10}$$

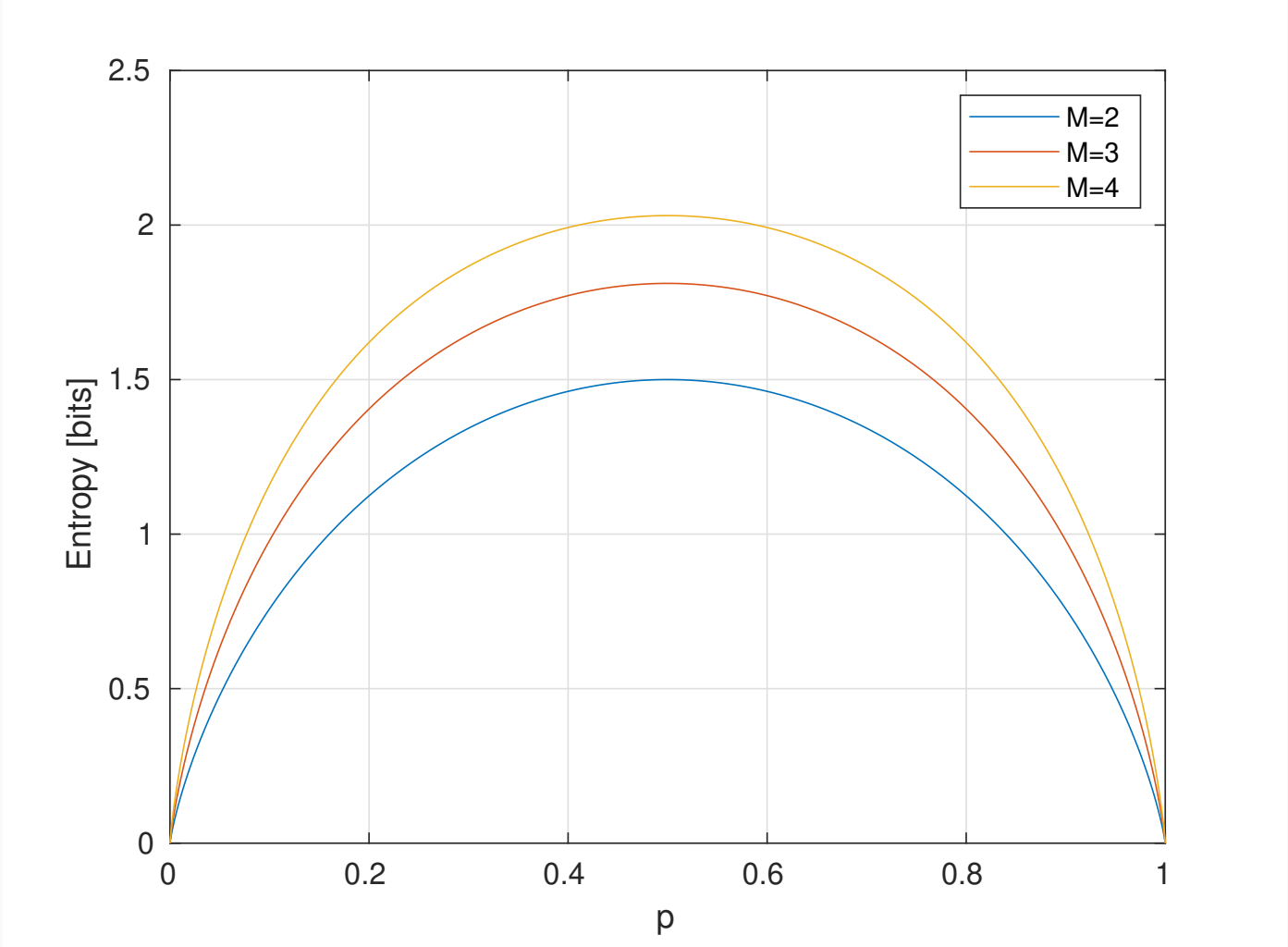
Example 1: Bernoulli distribution

Given a Bernoulli random variable X with parameter p we have that:

$$H(X) = -p\log(p) - (1-p)\log(1-p). \tag{11}$$



Example 2: Binomial distribution



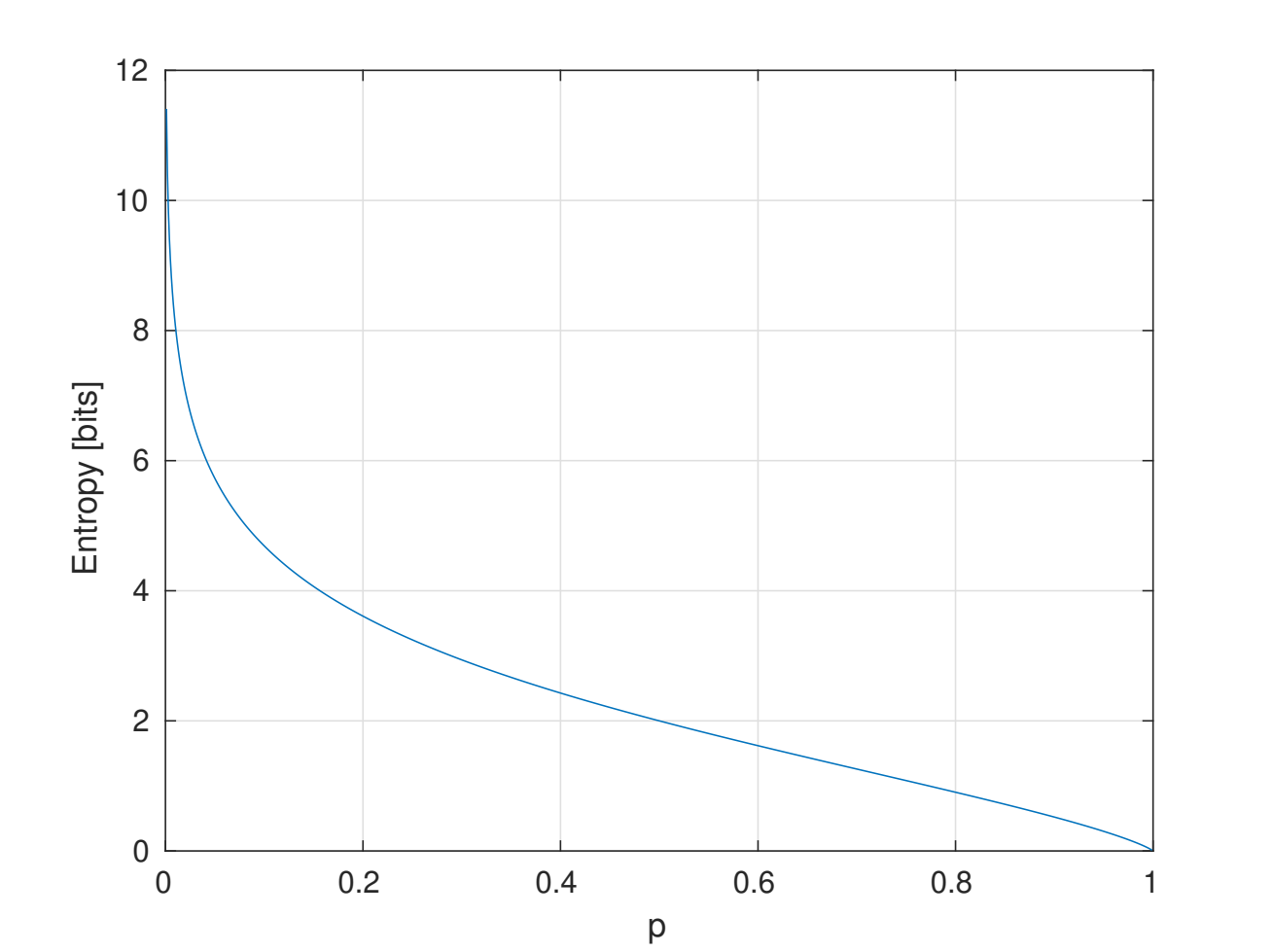
The value $p = 0.5$ results to maximizing the entropy for any value of M.

Example 3: Geometric distribution

Assuming $0 < p < 1$:

$$\begin{aligned} H(X) &= -\sum_{k=1}^{\infty} (1-p)^{k-1} p \log_2 \left((1-p)^{k-1} p \right) \\ &= -p \log_2 (p) \sum_{k=1}^{\infty} (1-p)^{k-1} - p \log_2 (1-p) \sum_{k=1}^{\infty} (k-1) (1-p)^{k-1} \\ &= \frac{-p \log_2 (p) - (1-p) \log_2 (1-p)}{p} \end{aligned} \tag{12}$$

Example 3: Geometric distribution



Example 3 continued

Intuition

The Geometric distribution gives us the probability mass function of the number of times that a binary experiment having two outcomes, $\{0, 1\}$, needs to be repeated in order to observe for the first time a value of 1. Assuming that at each repetition the probability of observing a value of 1 is p, and that repetitions result in independent outcomes, the Geometric distribution is obtained. As p becomes smaller, the probability that we need a large number of experiments in order to have a value of 1 becomes larger. As a result, more events start having non-negligible probability of occurrence, and the uncertainty increases.

Example 4: Uniform distribution

Let us consider a uniform random variable X taking one out of M possible values. The probability of occurrence of each one of the values is equal to $1/M$ and the entropy of X is equal to:

$$H(X) = -\sum_{i=1}^{M} \frac{1}{M} \log_2 \left(\frac{1}{M} \right) = log_2 M \tag{13}$$

The entropy is the same as the number of bits required for representing the values of X using a simple binary representation using the same number of bits for each one of the possible M values.

Joint entropy and conditional entropy: Studying multiple discrete random variables

Joint statistics of random variables

Definition: N-Dimensional random vectors

Let us consider a collection of sets $\mathcal{X}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,M_i}\}, i = 1, \ldots, N$, and a mapping that maps the outcome $s_k \in \mathcal{S}$ of an experiment to a point of the form $\left[x_1^{(k)}, \ldots, x_N^{(k)}\right]^T$, $x_i^{(k)} \in \mathcal{X}_i$, $i = 1, \ldots, N$. We then define the vector:

$$\mathbf{X} = \left[X_1(s_k), \ldots X_N(s_k)\right]^T, \tag{14}$$

produced by a mapping from $\mathcal{S}$ to the N-dimensional set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, as an N-dimensional random vector, or a vector of jointly distributed random variables $X_1, \ldots X_N$.

Definition: Joint probability mass function

We define as the joint probability mass function $p_{\mathbf{X}}[x_1, \ldots x_N]$ (or $p_{X_1,\ldots,X_N}[x_1, \ldots, x_N]$) the function that calculates the probability of the event $\Pr[X_1 = x_1, \ldots, X_N = x_N], i = 1, \ldots, N$.

From joint statistics to marginal and conditional statistics

Definition: Marginal PMF

Given the joint PMF of $X_1, \ldots, X_N$, we define the marginal PMF of X_k, $p_{X_k}[x]$, as the function calculating the probability $\Pr(X_k = x)$ that is calculated as:

$$p_{X_k}[x] = \Pr\left(X_k = x_k, \cap_{l=1, l\neq k}^{N} X_l \in \mathcal{X}_l\right) \tag{15}$$

Definition: Conditional PMF

For any given value $x \in \mathcal{X}_k$ we define the conditional PMF of X_k as:

$$p_{X_k|X_1=x^{(1)},\ldots X_{k-1}=x^{(k-1)},X_{k+1}=x^{(k+1)},\ldots,X_N=x^{(N)}}[x] =$$
$$\Pr\left[X_k = x | X_1 = x^{(1)}, \ldots X_{k-1} = x^{(k-1)}, X_{k+1} = x^{(k+1)}, \ldots, X_N = x^{(N)}\right]. \tag{16}$$

Expectation and conditional expectation

Definition: Expectation of a function of multiple random variables

Given a function $g(x_1, \ldots, x_N)$ of N random variables, we define its expectation as:

$$\mathbb{E}\{g(X_1, \ldots, X_N)\} = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_N \in \mathcal{X}_N} g(x_1, \ldots, x_N)\, p_{X_1,\ldots,X_N}[x_1, \ldots, x_N]. \tag{17}$$

Definition of conditional expectation

Given a function $g(x_1, \ldots, x_N)$ and knowledge that $X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_{k+1} = x_{k+1}, \ldots, X_N = x_N$, we define the conditional expectation as:

$$\mathbb{E}_{|X_1=x_1,\ldots,X_{k-1}=x_{k-1},X_{k+1}=x_{k+1},\ldots,X_N=x_N}\{g(X_1, \ldots, X_N)\} = \sum_{x_k \in \mathcal{X}_k} g(x_1, \ldots, x_n)\, p_{X_k|X_1=x_1,\ldots,X_{k-1}=x_{k-1},X_{k+1}=x_{k+1},\ldots,X_N=x_N}[x_k]. \tag{18}$$

Definition of joint and conditional entropy

Definition: Joint entropy

Given two discrete random variables X, Y and their joint distribution $p_{X,Y}[x, y]$ we define their joint entropy as:

$$H(X, Y) = -\mathbb{E}\left\{\log\left(p_{X,Y}[x, y]\right)\right\} = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x, y]\log\left(p_{X,Y}[x, y]\right). \tag{19}$$

Definition 2: Conditional entropy

Given two random variables X and Y and knowledge of the conditional distribution $p_{Y|X}[y|x]$, we define the conditional entropy as:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p_X[x]\, H(Y|X = x) \\ &= -\sum_{x\in\mathcal{X}} p_X[x] \sum_{y\in\mathcal{Y}} p_{Y|X=x}[y|x]\log\left(p_{Y|X=x}[y|x]\right) \\ &= -\mathbb{E}_{X,Y}\left\{\log\left(p_{Y|X}[Y|X]\right)\right\}. \end{aligned} \tag{20}$$

Example

Let X and Y be two independent random variables. We can then calculate their joint entropy as:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \log\left(p_{X,Y}[x,y]\right) p_{X,Y}[x,y] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \left(\log\left(p_X[x]\right)+\log\left(p_Y[y]\right)\right) p_X[x]\, p_Y[y] \\
&= -\sum_{x\in\mathcal{X}} p_X[x]\log\left(p_X[x]\right)\sum_{y\in\mathcal{Y}} p_Y[y] \\
&\quad -\sum_{y\in\mathcal{Y}} p_Y[y]\log\left(p_Y[y]\right)\sum_{x\in\mathcal{X}} p_X[x] \\
&= H(X)+H(Y).
\end{aligned}
\tag{21}
$$

Connection between join and conditional entropy

Chain rule theorem

$$H(X,Y) = H(X) + H(Y|X) \tag{22}$$

Proof:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(p_X[x]\,p_{Y|X}[y|x]\right) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(p_X[x]\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(p_{Y|X}[y|x]\right) \\ &= H(X) + H(Y|X)\,. \end{aligned} \tag{23}$$

Extending the chain rule

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

Relative entropy and mutual information

Relative entropy $D(p_X||q_X)$

- It measures the distance between two distributions $p_X[x]$ and $q_X[x]$ and as such it measures the inefficiency resulting from the assumption that X is distributed according to $q_X[x]$ when it is actually distributed according to $p_X[x]$.
- If X follows a distribution $p_X[x]$ we can then construct a code with average description length $H(p_X)$. However, if for the same random variable we use a code that is constructed for a random variable of distribution $q_X[x]$ we would then need an average length of $H(p_X) + D(p_X||q_X)$ bits to describe the random variable.

Formal definition of relative entropy

Definition: Relative entropy/Kullback-Leibler distance

Definition: The relative entropy (also called Kullback-Leibler distance) between two probability density functions $p_X[x]$ and $q_X[x]$ is given as:

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

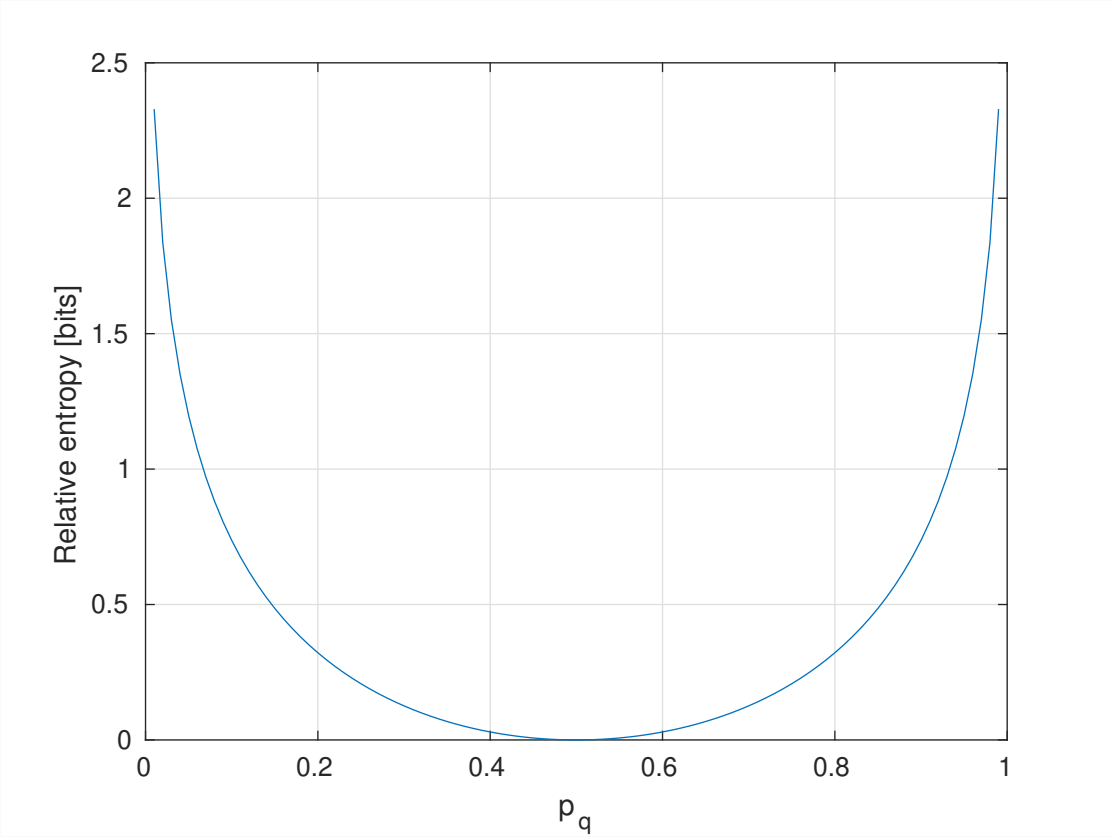
Remarks

1. If $q_X[x]$ takes a value of zero at points for which $p_X[x]$ is non zero, then $D(p_X||q_X)$ takes a value of infinity.
2. Relative entropy is not a true distance metric since it is not symmetric, i.e. $D(p_X||q_X) \neq D(q_X||p_X)$.

Example

Let X be a Bernoulli variable having a true value of parameter p equal to $1/2$. Let $q_X(x)$ be also a Bernoulli distribution, having a p parameter equal to p_q. We then have that:

$$D(p_X||q_X) = -\frac{1}{2}log_2(2p_q) - \frac{1}{2}\log_2(2(1-p_q)) \tag{24}$$



Mutual information

Definition: Mutual information

Let X and Y be two random variables characterized by a joint probability mass function $p_{X,Y}[x,y]$ and by marginal probability mass functions $p_X[x]$ and $p_Y[y]$ respectively. We then define the mutual information $I(X;Y)$ as the relative entropy between the joint distribution $p_{X,Y}[x,y]$ and the product distribution $p_X[x]\,p_Y[y]$, i.e., the quantity:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(\frac{p_{X,Y}[x,y]}{p_X[x]\,p_Y[y]}\right)$$

Remark

Essentially, the mutual information measures the distance between the joint distribution of X and Y and the joint distribution that these two variables would have in case that they were independent.

From entropy to mutual information

Linking entropy and mutual information

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(\frac{p_{X,Y}[x,y]}{p_X[x]\,p_Y[y]}\right)\\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\frac{p_X[x|y]}{p_X[x]}\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(p_X[x]\right) - \left[-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}[x,y]\log\left(p_X[x|y]\right)\right]\\
&= H(X) - H(X|Y)\,.
\end{aligned}
\tag{25}
$$

Some properties of mutual information

List of important properties

1. $I(X;Y) = H(X) - H(X|Y)$
2. $I(X;Y) = H(Y) - H(Y|X)$
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
4. $I(X;Y) = I(Y;X)$
5. $I(X;X) = H(X)$

Chain rules for Entropy/Mutual Information calculation

Chain rule for entropy

Entropy chain rule theorem

Let $X_1, \ldots, X_n$ be n discrete random variables following a joint probability mass functions $p[x_1, \ldots, x_n]$. We then have that:

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

Proof: By applying repeatedly the chain rule proven earlier for the two variable case, we obtain:

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

Conditional mutual information and information chain rule

Definition: Conditional mutual information and information chain rule

Let X, Y and Z be random variables described by their joint probability mass function. The conditional mutual information of X and Y given Z is defined as:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}_{X,Y,Z}\left\{\log\left(\frac{p[X,Y|Z]}{p[X|Z]\,p[Y|Z]}\right)\right\}.$$

Mutual information chain rule

Theorem: For mutual information, the following chain rule holds

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1).$$

Proof: $I(X_1, X_2, \ldots, X_n; Y) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n, |Y)$

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

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

Some useful inequalities in information theory

Definition of convex and concave functions

Definition: Convex function

A function $f(x)$ is called convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and any $\lambda \in [0, 1]$, property:

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

holds. In case that equality holds only for $\lambda = 0$, or $\lambda = 1$, function $f(x)$ is called strictly convex.

Definition: Concave function

A function $f(x)$ is called concave if $-f(x)$ is convex.

Theorem: Convexity and second derivative

If a function $f(x)$ has a non-negative second derivative over an interval, then it is convex over that interval. If the second derivative is positive, then the function is strictly convex.

Jensen's inequality

Theorem: Jensen inequality

For a convex function $f(X)$ of a random variable X we have that:

$$\mathbb{E}\{f(X)\} \geq f(\mathbb{E}\{X\}).$$

Proof: We focus on discrete random variables and apply induction with respect to the number of mass points of X. That is, we start by initially assuming a random variable X with two mass points x_1 and x_2, and probability mass function $\Pr(X = x_i) = p_i, i = 1, 2$. In order then for the Jensen inequality to hold we must have that:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \tag{26}$$

which is always true since function $f(X)$ is convex.

Jensen's inequality 2/2

Proof continued

Let us now assume that Jensen's inequality holds for a random variable with a distribution with $k-1$ mass points and try to extend it to a random variable with k mass points, having a probability mass function $\Pr(X = x_i) = p_j, j = 1, \ldots, k$. In such a case, by introducing $p_i' = p_i/(1-p_k), i = 1, 2, \ldots, k-1$ we have that:

$$\begin{aligned}\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \\ &\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \\ &\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) = f\left(\sum_{i=1}^{k} p_i x_i\right).\end{aligned}$$

Consequences of Jensen's inequality

Theory: Non negativity of differential entropy

Let $p_X[x], q_X[x], x \in \mathcal{X}$, be two probability mass functions. Then:

$$D(p_X||q_X) \geq 0,$$

with equality holding only if $p[x] = q[x]$ for all x.

Proof: Let us define as A the subset of $\mathcal{X}$ for which $A = \{x : p[x] > 0\}$. We then have that:

$$-D(p||q) = -\sum_{x \in A} p[x] \log\left(\frac{p[x]}{q[x]}\right) = \sum_{x \in A} p[x] \log\left(\frac{q[x]}{p[x]}\right).$$

By exploiting the concavity of logarithm and Jensen's inequality, we then obtain:

$$-D(p||q) \leq \log\left(\sum_{x \in A} p[x] \frac{q[x]}{p[x]}\right) = \log\left(\sum_{x \in A} q[x]\right) \leq \log 1 = 0.$$

Consequences of Jensen's inequality

Proof continued

Note that in the above procedure we have used the inequalities:

$$-D\left(p||q\right)=\sum_{x\in A}p\left[x\right]\log\left(\frac{q\left[x\right]}{p\left[x\right]}\right)\leq\log\left(\sum_{x\in A}p\left[x\right]\frac{q\left[x\right]}{p\left[x\right]}\right).$$

and

$$\log\left(\sum_{x\in A}p\left[x\right]\frac{q\left[x\right]}{p\left[x\right]}\right)\leq\log\left(\sum_{x\in\mathcal{X}}p\left[x\right]\frac{q\left[x\right]}{p\left[x\right]}\right)$$

Note however that the first inequality becomes a strict equality if and only if $q\left[x\right]=cp\left[x\right]$. We then have that $\sum_{x\in\mathcal{X}}q\left[x\right]=\sum_{x\in\mathcal{X}}p\left[x\right]\frac{q[x]}{p[x]}=c$. Moreover, we also have that $\sum_{x\in A}q\left[x\right]=\sum_{x\in\mathcal{X}}q\left[x\right]$ only if $c=1$. That proves the second part of the theorem.

Consequences of Jensen's inequality

Theorem: Non negativity of mutual information

For any two random variables X, Y, we have that $I(X;Y) \geq 0$ where equality holds if and only if X and Y are independent.

Proof: Since $I(X;Y) = D(p_{X,Y}[x,y] \,||\, p_X[x]\, p_Y[y])$, based on the previous theorem we obtain the result of this theorem.

Note: We can similarly prove that $I(X;Y|Z) \geq 0$.

Entropy related inequalities 1/2

Property: Bounding the entropy

Let X be a discrete random variable having an entropy equal to $H(X)$. Let $\mathcal{X}$ be the set of possible values of X and $|\mathcal{X}|$ its cardinality. We then have that $H(X) \leq \log|\mathcal{X}|$, with equality holding only for the case of a uniform distribution over $\mathcal{X}$.

Proof: Let $p_X[x]$ be the probability mass function of X and let $u[x] = \frac{1}{|\mathcal{X}|}$ be a uniform, over the possible values of X, distribution. By calculating the Kullback-Leibler distance of $p_X[x]$ and $u[x]$ we have that:

$$D(p_X||u) = \sum_{x\in\mathcal{X}} p_X[x] \log\left(\frac{p_X[x]}{u[x]}\right) = \log|X| - H(X).$$

As a result, due to non-negativity of Kullback-Leibler distance, we obtain that $H(X) \leq \log|\mathcal{X}|$.

Entropy related inequalities 2/2

Property: Bounding conditional entropy

Let X, Y be two random variables. It then holds that:

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

with equality holding if and only if X and Y are independent.

Proof: As we have seen earlier, mutual information $I(X;Y)$, defined as $I(X;Y) = H(X) - H(X|Y)$, is non negative. Moreover, as we have seen earlier, the mutual information equals zero if and only if X and Y are independent. As a result, it holds that $H(X) = H(X|Y)$, if and only if X and Y are independent.

Intuitively, this inequality implies that in the presence of knowledge for random variable Y, the average uncertainty concerning random variable X reduces.

Compression and it's connection to entropy: The Asymptotic Equipartition Property

Introduction

- We focus on sequences $\{X_1, \ldots, X_n\}$ of i.i.d. random variables and their properties.
- Using the law of large numbers, the asymptotic equipartition property states that the sample entropy:

$$\frac{1}{n} \log \frac{1}{p\left[X_1, X_2, \ldots, X_n\right]} \tag{27}$$

  converges to the actual entropy $H(X)$ of the distribution of these samples.
- Equivalently, this allows us to state that for a sequence $\{X_1, \ldots, X_n\}$, the probability $p\left[X_1, X_2, \ldots, X_n\right]$ converges to a value $2^{-nH(X)}$.
- We can then separate all sequences in two sets, a set of **typical sequences** that approximately satisfy this property, and the set of **nontypical sequences** that do not.

Reminder: Types of convergence of random variables

Convergence in probability

We say that a sequence $\{X_1, \ldots, X_n\}$ of random variables converges in probability to a random variable X if for every ϵ,
$\Pr\{|X_n - X| > \epsilon\} \to 0$.

Convergence in mean square

We say that a sequence $\{X_1, \ldots, X_n\}$ of random variables converges in the mean square error sense to a random variable X, if
$\mathbb{E}\left\{(X - X_n)^2\right\} \to 0$.

Convergence with probability 1

We say that a sequence $\{X_1, \ldots, X_n\}$ of random variables converges with probability one to a random variable X, if
$\Pr\{\lim_{n\to\infty} X_n = X\} = 1$.

Asymptotic equipartition property

The asymptotic equipartition property theorem

If random variables $X_1, X_2, \ldots$ are i.i.d. following a probability mass function $p_X[x]$, then:

$$-\frac{1}{n}\log\left(p\left[X_1,\ldots,X_n\right]\right) \to H(X) \tag{28}$$

in probability.

Proof: Since $X_1,\ldots,X_n$ are i.i.d., also random variables $\log p[X_1],\ldots,\log p[X_n]$ are i.i.d.. As a result, by applying the weak law of large numbers we have that (considering convergence in probability):

$$-\frac{1}{n}\log\left(p\left[X_1,\ldots,X_n\right]\right) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left[X_i\right]\right) \to \mathbb{E}\left\{\log\left(p\left[X\right]\right)\right\} = H(X). \tag{29}$$

Typical sets

Definition: Typical set

Let $\mathcal{X}^n$ denote the set of all the sequences of the form $\{x_1, \ldots, X_n\}$. Given a value ϵ, we define as the typical set $A_\epsilon^{(n)}$ (with respect to the probability mass function $p_X[x]$) the set of sequences $\{x_1, x_2, \ldots, x_n\} \in \mathcal{X}^n$ that satisfy the property:

$$2^{-n(H(X)+\varepsilon)} \leq p[x_1, x_2, \ldots, x_n] \leq 2^{-n(H(X)-\varepsilon)} \tag{30}$$

Properties of the typical set

Theorem: Properties of the typical set

1. If $\{x_1, x_2, \ldots, x_n\} \in A_\epsilon^{(n)}$, then
   $H(X) - \epsilon \leq -\frac{1}{n} \log\left(p[x_1, \ldots, x_n]\right) \leq H(X) + \epsilon$.
2. $\Pr\left\{A_\epsilon^{(n)}\right\} > 1 - \epsilon$ for n sufficiently large.
3. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$.
4. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)\, 2^{n(H(X)-\epsilon)}$, for n sufficiently large.

Proof:

- Property 1 follows from the definition of the typical set.

Properties of the typical set

Proof continued

- Based on the asymptotic equipartition theorem and the definition of convergence in probability, we have that for any $\delta > 0$, there exists an n_0 such that for all $n \geq n_0$, we have that:

$$\Pr\left\{\left|-\frac{1}{n}\log p\left[X_1, X_2, \ldots, X_n\right] - H\left(X\right)\right| < \epsilon\right\} > 1 - \delta. \tag{31}$$

  By selecting $\delta = \epsilon$ we then obtain Property 2.

- To prove Property 3 we note that:

$$1 = \sum_{\mathbf{x}\in\mathcal{X}^n} p\left[\mathbf{x}\right] \geq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p\left[\mathbf{x}\right] \geq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)}\left|A_\epsilon^{(n)}\right|. \tag{32}$$

- To prove Property 4 we note that:

$$1 - \epsilon < \Pr\left\{A_\epsilon^{(n)}\right\} \leq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)}\left|A_\epsilon^{(n)}\right| \tag{33}$$

Consequences of the asymptotic equipartition property

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. random variables having a mass function $p[x]$. Letalso $\epsilon > 0$. Then there exists a code that maps each sequence X^n of length n into a binary string of length $l(X^n)$, such that the mapping is one-to-one and

$$\mathbb{E}\left\{\frac{1}{n} l(X^n)\right\} \leq H(X) + \epsilon, \tag{34}$$

for sufficiently large n.

Proof: We start by dividing all sequences of n symbols into the typical set $A_\epsilon^{(n)}$ and its complement. Since the cardinality of the typical set is at most $2^{n(H(X)+\epsilon)}$, we would need at most $n(H(X)+\epsilon)+1$ bits, if we were to encode all elements of $A_\epsilon^{(n)}$ using the same number of bits . On the other hand, for the non typical sequences we could use at most a binary mapping of length $n \log |X| + 1$ to encode them.

Consequences of the asymptotic equipartition property

Proof continued

Hence, using the above coding scheme and an additional bit to indicate whether a sequence belongs in the typical or atypical set; we have that the expected length of a code word is:

$$\begin{aligned}
\mathbb{E}\left\{l\left(X^n\right)\right\} &= \sum_{x^n} p\left[x^n\right] l\left(x^n\right) \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p\left[x^n\right] l\left(x^n\right) + \sum_{x^n \notin A_\epsilon^{(n)}} p\left[x^n\right] l\left(x^n\right) \\
&\leq \Pr\left\{x^n \in A_\epsilon^{(n)}\right\}\left(n\left(H\left(X\right)+\epsilon+2\right)\right) + \Pr\left\{x^n \notin A_\epsilon^{(n)}\right\}\left(n\log\left|\mathcal{X}\right|+2\right) \\
&\leq n\left(H\left(X\right)+\epsilon\right) + \epsilon n\left(\log\left|\mathcal{X}\right|\right) + 2 = n\left(H\left(X\right)+\epsilon'\right)
\end{aligned} \tag{35}$$

where $\epsilon' = \epsilon + \epsilon\log\left|\mathcal{X}\right| + \frac{2}{n}$, can take arbitrarily small values by appropriately choosing n.