

**Proposition 1 (Information inequality).** *Let  $p(x), q(x), x \in \mathcal{X}$  be two probability mass functions. Then*

$$D(p||q) \geq 0$$

*with equality if and only if*

$$p(x) = q(x) \quad \text{for all } x.$$

*Proof.* Let  $A = \{x : p(x) > 0\}$  be the support set of  $p(x)$ . Then

$$\begin{aligned} D(p||q) &= \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in A} p(x) \left[ -\log \frac{q(x)}{p(x)} \right] \\ &\geq -\log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && (1) \\ &= -\log \sum_{x \in A} q(x) \geq -\log \sum_{x \in \mathcal{X}} q(x) \\ &= -\log 1 = 0 \end{aligned}$$

where (1) follows by Jensen's inequality. Since  $-\log t$  is a strictly concave function of  $t$ , we have equality in (1) if and only if  $q(x)/p(x) = 1$  everywhere. Hence we have  $D(p||q) = 0$  if and only if  $p(x) = q(x)$  for all  $x$ .  $\square$

**Corollary 1 (Non-negativity of mutual information).** *For any two random variables,  $X, Y$ ,*

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

*with equality if and only if  $X$  and  $Y$  are independent.*

*Proof.*  $I(X; Y) = D(p_{X,Y}(x, y) \| p_X(x)p_Y(y)) \geq 0$ , with equality if and only if  $p(x, y) = p(x)p(y)$ , i.e.,  $X$  and  $Y$  are independent. □

**Proposition 2.**  $H(X) \leq \log |\mathcal{X}|$ , where  $|\mathcal{X}|$  denotes the number of elements in the range of  $X$ , with equality if and only if  $X$  has a uniform distribution over  $\mathcal{X}$ .

*Proof.* Let  $u(x) = \frac{1}{|\mathcal{X}|}$  be the uniform probability mass function over  $\mathcal{X}$ , and let  $p(x)$  be the probability mass function for  $X$ . Then

$$D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X).$$

Hence by the non-negativity of relative entropy,

$$0 \leq D(p||u) = \log |\mathcal{X}| - H(X).$$

□

**Proposition 3 (Conditioning reduces entropy).**

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

*with equality if and only if  $X$  and  $Y$  are independent.*

*Proof.*  $0 \leq I(X; Y) = H(X) - H(X|Y).$

□

**Proposition 4 (Independence bound on entropy).** *Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then*

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

*with equality if and only if the  $X_i$  are independent.*

*Proof.* By the chain rule for entropies

$$\begin{aligned} H(X_1, \dots, X_n) &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \\ &\leq \sum_{i=1}^n H(X_i) \end{aligned}$$

where the inequality follows directly from the previous theorem. We have equality if and only if  $X_i$  is independent of  $X_{i-1}, \dots, X_1$  for all  $i$ , i.e., if and only if the  $X_i$ 's are independent. □

## The log sum inequality and its applications

**Proposition 5 (Log sum inequality).** For non-negative numbers,  $a_1, a_2, \dots, a_n$  and  $b_1, b_2, \dots, b_n$ ,

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

with equality if and only if  $\frac{a_i}{b_i} = \text{const.}$

**Exercise:** Deduce this inequality from Jensen's inequality and the fact that  $t \mapsto t \log t$  is a strictly convex function.

**Proposition 6.**  $D(p\|q)$  is convex in the pair  $(p, q)$ , i.e., if  $(p_1, q_1)$  and  $(p_2, q_2)$  are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1 - \lambda)p_2\|\lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1\|q_1) + (1 - \lambda)D(p_2\|q_2)$$

for all  $0 \leq \lambda \leq 1$ .

*Proof.* We apply the log sum inequality to a term on the left hand side, i.e.,

$$\begin{aligned} (\lambda p_1(x) + (1 - \lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \\ \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda p_2(x)} + (1 - \lambda)p_2(x) \frac{(1 - \lambda)p_1(x)}{(1 - \lambda)p_2(x)} \end{aligned} \quad (2)$$

Summing this over all  $x$ , we obtain the desired property.  $\square$

**Proposition 7 (Concavity of entropy).**  $H(p)$  is a concave function of  $p$ .

*Proof.*

$$H(p) = \log |\mathcal{X}| - D(p||u),$$

where  $u$  is the uniform distribution on  $|\mathcal{X}|$  outcomes. The concavity of  $H$  then follows directly from the convexity of  $D$ . □

### Data processing inequality

**Definition 1.** Random variables  $X, Y, Z$  are said to form a Markov chain in that order (denoted by  $X \rightarrow Y \rightarrow Z$ ) if the conditional distribution of  $Z$  given  $X, Y$  is just the conditional distribution of  $Z$  given  $Y$ . Specifically,  $X, Y$  and  $Z$  form a Markov chain  $X \rightarrow Y \rightarrow Z$  if the joint probability mass function can be written as

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y).$$

Some simple consequences are as follows:

- $X \rightarrow Y \rightarrow Z$  if and only if  $X$  and  $Z$  are conditionally independent given  $Y$ .

The Markovian property implies conditional independence because

$$\begin{aligned} p_{X,Z|Y}(x, z|y) &= \frac{p_{X,Y,Z}(x, y, z)}{p_Y(y)} \\ &= \frac{p_{X,Y}(x, y)p_{Z|Y}(z|y)}{p_Y(y)} = p_{X|Y}(x|y)p_{Z|Y}(z|y). \end{aligned}$$

This is the characterization of Markov chains that can be extended to define Markov fields, which are  $n$ -dimensional random processes in which the interior and exterior are independent given the values on the boundary.

- $X \rightarrow Y \rightarrow Z$  implies that  $Z \rightarrow Y \rightarrow X$ . Thus the condition is sometimes written  $X \leftrightarrow Y \leftrightarrow Z$ ,
- If  $Z = f(Y)$  then  $X \rightarrow Y \rightarrow Z$ .

**Proposition 8 (Data processing inequality).** *If  $X \rightarrow Y \rightarrow Z$ , then  $I(X; Y) \geq I(X; Z)$ .*

*Proof.* By the chain rule, we can expand mutual information in two different ways.

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y) \end{aligned}$$

Since  $X$  and  $Z$  are conditionally independent given  $Y$ , we have  $I(X; Z|Y) = 0$ . Since  $I(X; Y|Z) \geq 0$ , we have

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

We have equality if and only if  $I(X; Y|Z) = 0$ , i.e.,  $X \rightarrow Z \rightarrow Y$  forms a Markov chain. Similarly, one can prove that  $I(Y; Z) \geq I(X; Z)$ . □

**Corollary 2.** *In particular, if  $Z = g(Y)$ , we have  $I(X; Y) \geq I(X; g(Y))$ .*

### Sufficient statistics

Suppose we have a family of probability mass functions  $\{f_\theta(x)\}$  indexed by  $\theta$ , and let  $X$  be a sample from a distribution in this family. Let  $T(X)$  be any statistic (function of the sample) like the sample mean or sample variance. Then  $\theta \rightarrow X \rightarrow T(X)$ , and by the data processing inequality, we have

$$I(\theta; T(X)) \leq I(\theta; X)$$

for any distribution on  $\theta$ . However, if equality holds, no information is lost.

A statistic  $T(X)$  is called sufficient for  $\theta$  if it contains all the information in  $X$  about  $\theta$ .

**Definition 2.** A function  $T(X)$  is said to be a **sufficient statistic** relative to the family  $\{f_\theta\}$  if  $X$  is independent of  $\theta$  given  $T(X)$ , i.e.,  $\theta \rightarrow T(X) \rightarrow X$  forms a Markov chain.

This is the same as the condition for equality in the data processing inequality.

$$I(\theta; X) = I(\theta; T(X))$$

for all distributions on  $\theta$ . Hence sufficient statistics preserve mutual information and conversely.

*Example 1.* Let  $X_1, X_2, \dots, X_n, X_i \in \{0, 1\}$ , be an independent and identically distributed sequence of coin tosses of a coin with unknown parameter  $\theta = \mathbb{P}\{X_i = 1\}$ . Given  $n$ , the number of 1's is a sufficient statistic for  $\theta$ . Here  $T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i$ . In fact, we can show that given  $T$ , all sequences having that many 1's are equally likely and independent of the parameter  $\theta$ . Specifically,

$$\mathbb{P} \left\{ (X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n) \mid \sum_{i=1}^n X_i = k \right\} = \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } \sum x_i = k, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Thus  $\theta \rightarrow \sum X_i \rightarrow (X_1, X_2, \dots, X_n)$  forms a Markov chain, and  $T$  is a sufficient statistic for  $\theta$ .

## The Asymptotic Equipartition Property

**Proposition 9 (AEP).** *If  $X_1, X_2, \dots$  are i.i.d.  $\sim p$ , then*

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X), \quad \text{almost surely.}$$

*Proof.* Functions of independent random variables are also independent random variables. Thus, since  $X_i$  are i.i.d., so are  $\log p(X_i)$ . Hence by the strong law of large numbers,

$$\begin{aligned} -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) &= -\frac{1}{n} \sum_i \log p(X_i) \\ &\rightarrow -\mathbb{E} \log p(X) \quad \text{almost surely} \\ &= H(X), \end{aligned}$$

which proves the theorem. □

**Definition 3.** The **typical** set  $A_\epsilon^n$  with respect to  $p(x)$  is the set of sequences  $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$  with the following property:

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

**Proposition 10.** (a) If  $(x_1, x_2, \dots, x_n) \in A_\epsilon^n$ , then

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots, x_n) \leq H(X) + \epsilon.$$

(b)  $\mathbb{P} \{A_\epsilon^n\} > 1 - \epsilon$  for  $n$  sufficiently large.

(c)  $|A_\epsilon^n| \leq 2^{n(H(X)+\epsilon)}$ , where  $|A|$  denotes the number of elements in the set  $A$ .

(d)  $|A_\epsilon^n| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$  for  $n$  sufficiently large.

Thus the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly  $2^{nH}$ .

*Proof.* The proof of property (a) is immediate from the definition of  $A_\epsilon^n$ . Property (b) follows directly from (AEP), since the probability of the event  $(X_1, X_2, \dots, X_n) \in A_\epsilon^n$  tends to 1 as  $n \rightarrow \infty$ . Thus for any  $\delta > 0$ , there exists an  $N$ , such that for all  $n \geq N$ , we have

$$\mathbb{P} \left\{ \left| -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) - H(X) \right| < \epsilon \right\} > 1 - \delta.$$

Setting  $\delta = \epsilon$  we obtain the second part of the theorem.

To prove property (c), we write

$$\begin{aligned} 1 &= \sum_{x \in \mathcal{X}^n} p(x) \\ &\geq \sum_{x \in A_\epsilon^n} p(x) \\ &\geq \sum_{x \in A_\epsilon^n} 2^{-n(H(X)+\epsilon)} \\ &= 2^{-n(H(X)+\epsilon)} |A_\epsilon^n| \end{aligned}$$

where the second inequality follows from the definition. Hence

$$|A_\epsilon^n| \leq 2^{n(H(X)+\epsilon)}.$$

Property (d) is similar.

□

## Data Compression

Let  $X_1, X_2, \dots, X_n$  be i.i.d. random variables drawn from the probability mass function  $p(x)$ . We wish to find short descriptions for such sequences of random variables. We divide all sequences in  $\mathcal{X}^n$  into two sets: the typical set  $A_\epsilon^n$  and its complement.

We order all elements in each set according to some order (say lexicographic).

Then we can represent each sequence of  $A_\epsilon^n$  by giving the index of the sequence in the set. Since there are  $\leq 2^{n(H+\epsilon)}$  sequences in  $A_\epsilon^n$ , the indexing requires no more than  $n(H + \epsilon) + 1$  bits. (The extra bit may be necessary because  $n(H + \epsilon)$  may not be an integer.) We prefix all these sequences by a 0, giving a total length of  $\leq n(H + \epsilon) + 2$  bits to represent each sequence in  $A_\epsilon^n$ .

Similarly, we can index each sequence not in  $A_\epsilon^n$  by using not more than  $n \log |\mathcal{X}| + 1$  bits. Prefixing these indices by 1, we have a code for all the sequences in  $\mathcal{X}^n$ .

Note the following features of the above coding scheme.

- The code is one-to-one and easily decodable. The initial bit acts as a flag bit to indicate the length of the codeword that follows.
- We have used a brute force enumeration of the atypical set  $(A_\epsilon^n)^c$ , without taking into account the fact that the number of elements in  $(A_\epsilon^n)^c$  is less than the number of elements in  $\mathcal{X}^n$ . Surprisingly, this is good enough to yield an efficient description.
- The typical sequences have short descriptions of length  $\approx nH$ .

We will use the notation  $x^n$  to denote a sequence  $x_1, x_2, \dots, x_n$ . Let  $l(x^n)$  be the length of the codeword corresponding to  $x^n$ . If  $n$  is sufficiently large so that  $\mathbb{P}\{A_\epsilon^n\} \geq 1 - \epsilon$ , then the expected length of the codeword is

$$\begin{aligned}
\mathbb{E}(l(X^n)) &= \sum_{x^n} p(x^n)l(x^n) = \sum_{x^n \in A_\epsilon^n} p(x^n)l(x^n) + \sum_{x^n \in (A_\epsilon^n)^c} p(x^n)l(x^n) \\
&\leq \sum_{x^n \in A_\epsilon^n} p(x^n)[n(H + \epsilon) + 2] + \sum_{x^n \in (A_\epsilon^n)^c} p(x^n)[n \log |\mathcal{X}| + 2] \\
&= \mathbb{P}\{A_\epsilon^n\} [n(H + \epsilon)] + \mathbb{P}\{(A_\epsilon^n)^c\} [n \log |\mathcal{X}| + 2] \\
&\leq n(H + \epsilon) + \epsilon n(\log |\mathcal{X}|) + 2 \\
&= n(H + \epsilon'),
\end{aligned}$$

where  $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + 2/n$  can be made arbitrarily small by an appropriate choice of  $\epsilon$  followed by an appropriate choice of  $n$ . Hence we have proved the following theorem.

**Proposition 11.** *Let  $X^n$  be i.i.d.  $\sim p(x)$ . Let  $\epsilon > 0$ . Then there exists a code which maps sequences  $x^n$  of length  $n$  into binary strings such that the mapping is one-to-one (and therefore invertible) and*

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

*for  $n$  sufficiently large.*

Thus we can represent sequences  $X^n$  using  $nH(X)$  bits on average.