INFORMATION THEORY & CODING

Part 2 : Entropy

Dr. Rui Wang

Department of Electrical and Electronic Engineering

Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

September 19, 2023

Outline

- Most of the basic definitions will be introduced.

  Entropy, Joint Entropy, Conditional Entropy, Relative Entropy, Mutual Information, etc.

- *Relationships* among basic definitions.

  Chain Rules.

What is *Information*?

- To have a quantitative measure of information contained in an event, we consider intuitively the following:
  - Information contained in events should be defined in terms of the *uncertainty/probability* of the events.
  - *Monotonous: Less certain (small probability)* events should contain *more information*.
  - *Additive:* The total information of *unrelated/independent* events should equal the *sum* of the information of each individual event.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Information Measure of Random Events

A *natural* measure of the *uncertainty* of an event A is the probability $Pr(A)$ of A.

To satisfy the *monotonous* and *additive* properties, the information in the event A could be defined as

$$I(A)_{\text{self-info}} = -\log \Pr(A).$$

If $\Pr(A) > \Pr(B)$, then $I(A) < I(B)$.

If A, B are independent, then $I(A + B) = I(A) + I(B)$.

Information Unit

$$\log_2 \quad : \quad \text{bit}$$
$$\log_e \quad : \quad \text{nat}$$
$$\log_{10} \quad : \quad \text{Hartley}$$

$$\log_a X = \frac{\log_b X}{\log_b a} = \log_a b \cdot \log_b X$$

Average Information Measure of a Discrete R.V.

$x_1, x_2, \cdots, x_q$: *Alphabet $\mathcal{X}$ (realizations) of discrete r.v. X*

$p_1, p_2, \cdots, p_q$: *Probability*

The *average* information of the r.v. X is

$$I(X) = \sum_{i=1}^{q} p_i \log(\frac{1}{p_i}),$$

where $\log \frac{1}{p_i}$ is the *self-information* of event $X = x_i$.

Entropy

Definition

The *entropy* of a discrete random variable X is given by

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= E\left[\log \frac{1}{p(X)}\right]. \end{aligned}$$

By convention, let $0 \log 0 = 0$ since $x \log x \to 0$ as $x \to 0$.

Lemma 2.1.1

$$H(X) \geq 0.$$

Proof.

Since $0 \leq p(x) \leq 1$, we have $\log \frac{1}{p(x)} \leq 0$. □

Lemma 2.1.2

$$H_b(X) = (\log_b a)H_a(X).$$

Proof.

Since $\log_b p = \log_b a \log_a p$. □

Entropy: An Example

Example 2.1.1

Let

$$X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p. \end{cases}$$

$$H(X) = -p \log p - (1-p)\log(1-p) = H(p).$$

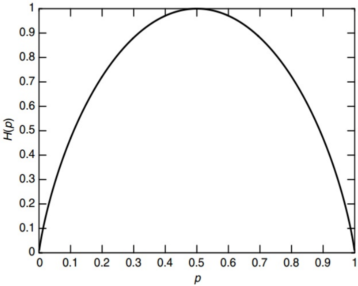


FIGURE 2.1. $H(p)$ vs. p.

Joint Entropy

Definition

The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is

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

If X and Y are *independent*, then

$$
\begin{aligned}
H(X, Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y) \log p(x)p(y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y) \log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x)p(y) \log p(y) \\
&= \sum_{y\in\mathcal{Y}} p(y)H(X) + \sum_{x\in\mathcal{X}} p(x)H(Y) \\
&= H(X) + H(Y).
\end{aligned}
$$

Conditional Entropy

- If $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y|X)$ is

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

Theorem 2.2.1 (Chain Rule)

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

The *joint entropy* of a pair of random variables = the entropy of one + the *conditional entropy* of the other.

$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x, y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Chain Rule

Theorem 2.2.1 (Chain Rule)

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

Corollary

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

Proof ?

Example

Example 2.2.1

Let (X, Y) have the following *joint distribution*:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

What are $H(X)$, $H(Y)$, $H(X, Y)$, $H(X|Y)$, and $H(Y|X)$?

$$H(X) = \frac{7}{4}\text{bits}, H(Y) = 2\text{bits}, H(X|Y) = \frac{11}{8}\text{bits},$$
$$H(Y|X) = \frac{13}{8}\text{bits}, H(X, Y) = \frac{27}{8}\text{bits}.$$

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Relative Entropy

- The *entropy* of a random variable is a measure of *the amount of information* required to describe the random variable.

- The *relative entropy* $D(p\|q)$ is a measure of *the distance between two distributions*. We need $H(p)$ bits on average to describe a random variable with distribution p, and need $H(p) + D(p\|q)$ bits on average to describe a random variable with distribution q from the distribution p point of view.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Relative Entropy

- The *relative entropy* or *Kullback-Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$
\begin{aligned}
D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&= E_p \log\frac{p(X)}{q(X)}.
\end{aligned}
$$

By convention, $0\log\frac{0}{0}=0$, $0\log\frac{0}{q}=0$ and $p\log\frac{p}{0}=\infty$.

Relative Entropy

- $D(p\|q) = D(q\|p)$?

Example 2.3.1

Let $\mathcal{X} = \{0, 1\}$ and consider two distributions p and q on $\mathcal{X}$. Let $p(0) = 1 - r$, $p(1) = r$, and let $q(0) = 1 - s$, $q(1) = s$. Then

$$D(p\|q) = (1 - r)\log\frac{1 - r}{1 - s} + r\log\frac{r}{s},$$
$$D(q\|p) = (1 - s)\log\frac{1 - s}{1 - r} + s\log\frac{s}{r}.$$

In general, $D(p\|q) \neq D(q\|p)$!

Mutual Information

Definition

Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)q(y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)\|p(x)p(y)) \\ &= E_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)} \end{aligned}$$

Relationships

Theorem 2.4.1 (Mutual information and entropy)

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

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

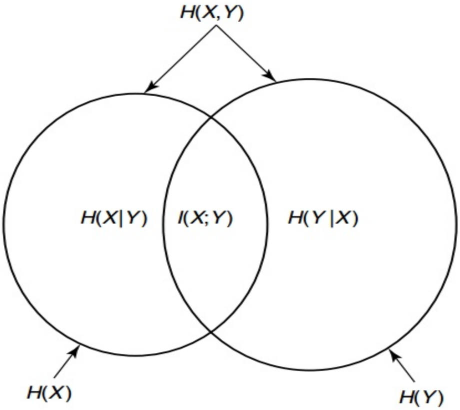
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = I(Y;X)$$

$$I(X;X) = H(X)$$

Relationships

- *Mutual information and entropy*



Chain Rules

Theorem 2.5.1 (Chain rule for entropy)

Let $X1, X2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

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

Proof.

Chain Rules

Definition

The *conditional mutual information* of random variable X and Y given Z is defined by

$$
\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\
&= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}
\end{aligned}
$$

Theorem 2.5.2 (Chain rule for mutual information)

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

Chain Rules

Definition

For joint probability mass functions $p(x, y)$ and $q(x, y)$, the *conditional relative entropy* $D(p(y|x)\|q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

$$D(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}$$

Chain Rules

Theorem 2.5.3 (Chain rule for relative entropy)

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x))$$

Proof.

$$\begin{aligned}
&D(p(x,y)\|q(x,y)) \\
&\quad = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&\quad = \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&\quad = \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&\quad = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x))
\end{aligned}$$

Reading & Homework

Reading: Chapter 2.1 - 2.5

Homework: Problems 2.1, 2.2, 2.3, 2.4