

Entropy, Relative Entropy, and Mutual Information

Some basic notions of Information Theory

Radu Trîmbițaș

October 2012

Outline

Contents

| | |
|--|-----------|
| 1 Entropy and its Properties | 1 |
| 1.1 Entropy | 1 |
| 1.2 Joint Entropy and Conditional Entropy | 3 |
| 1.3 Relative Entropy and Mutual Information | 5 |
| 1.4 Relationship between Entropy and Mutual Information | 6 |
| 1.5 Chain Rules for Entropy, Relative Entropy and Mutual information | 8 |
| 2 Inequalities in Information Theory | 10 |
| 2.1 Jensen inequality and its consequences | 10 |
| 2.2 Log sum inequality and its applications | 13 |
| 2.3 Data-processing inequality | 14 |
| 2.4 Sufficient statistics | 15 |
| 2.5 Fano's inequality | 16 |

1 Entropy and its Properties

1.1 Entropy

Entropy of a discrete RV

- a measure of uncertainty of a random variable
- X a discrete random variable

$$X \sim \left(\begin{matrix} x_i \\ p_i \end{matrix} \right)_{i \in I}, \mathcal{X} \text{ alphabet of } X, p(x) = P(X = x), \text{ mass function of } X$$

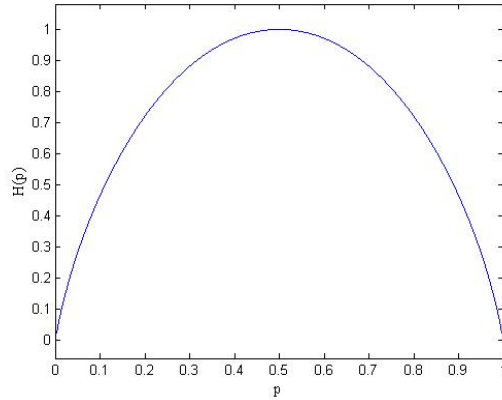


Figure 1: Graph of $H(p)$

Definition 1. The *entropy* of the discrete random variable X

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (1)$$

$$H(X) = E_p \left(\log \frac{1}{p(x)} \right) \quad \text{equivalent expression} \quad (2)$$

- measured in *bits*!
- base 2! $H_b(X)$ entropy in base b ; for $b = e$, measured in *nats*!
- convention $0 \log 0 = 0$, since $\lim_{x \searrow 0} x \log x = 0$

Entropy - Properties

Lemma 2. $H(X) \geq 0$

Lemma 3. $H_b(X) = \log_b a H_a(X)$

Example 4. Let the RV

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

$$H(X) = -p \log p - (1-p) \log(1-p) =: H(p) \quad (3)$$

$H(X) = 1$ bit when $p = \frac{1}{2}$. Graph in Figure 1

Entropy - Properties II

Entropy - Properties III

$$X : \begin{pmatrix} a & b & c & d \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

The entropy of X is

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = \frac{7}{4} \text{ bits}$$

Problem: Determine the value of X with the minimum number of binary questions.

Sol: Is $X = a$? Is $X = b$? Is $X = c$? The resulting expected number is $\frac{7}{4} = 1.75$ bits. See Lectures on Data Compression: the minimum expected number of binary question required to determine X lies between $H(X)$ and $H(X) + 1$.

1.2 Joint Entropy and Conditional Entropy

Joint Entropy and Conditional Entropy

- (X, Y) a pair of discrete RVs over the alphabets \mathcal{X}, \mathcal{Y}

$$X : \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}, \quad Y : \begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j \in J}$$

- *joint distribution* of X and Y

$$p(x, y) = P(X = x, Y = y), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

- *(marginal) distribution* of X

$$p_X(x) = p(x) = P(X = x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

- *(marginal) distribution* of Y

$$p_Y(y) = p(y) = P(Y = y) = \sum_{x \in \mathcal{X}} p(x, y)$$

Definition 5. The *joint entropy* $H(X, Y)$ of a pair of DRV $(X, Y) \sim p(x, y)$

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \quad (4)$$

also expressed as $H(X, Y) = -E(\log p(X, Y))$

Definition 6. $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \quad (5)$$

where

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

$$p(y|x) := P(Y = y|X = x) = \underbrace{\frac{P(Y = y, X = x)}{P(X = x)}}_{\text{conditional probability}} = \frac{p(x, y)}{p(x)}$$

- By computation

$$H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \quad (6)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \quad (7)$$

$$= -E(\log p(Y|X)) \quad (8)$$

- naturalness of last two definitions \leftarrow the entropy of a pair of RVs is the entropy of one plus the conditional entropy of the other – see next theorem

Theorem 7 (Chain Rule).

$$H(X, Y) = H(X) + H(Y|X). \quad (9)$$

Proof.

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

Equivalently (shorter proof): we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X)$$

and apply E to both sides. \square

Joint Entropy and Conditional Entropy - b

Corollary 8.

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

Example 9. Let (X, Y) have the joint distribution

| $Y \setminus X$ | 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 |

- marginal distributions $X: (\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{8})$; $Y: (\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4})$
- $H(X) = \frac{7}{4}$ bits, $H(Y) = 2$ bits
- conditional entropy

$$\begin{aligned} H(X|Y) &= \sum_{i=1}^4 p(Y=i) H(X|Y=i) \\ &= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) \\ &\quad + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\ &= \frac{1}{4} \cdot \left(\frac{7}{4} + \frac{7}{4} + 2 + 0\right) = \frac{11}{8} \text{bits} \end{aligned}$$

- $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.
- **Remark.** If $H(X) \neq H(Y)$ then $H(Y|X) \neq H(X|Y)$. However $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

1.3 Relative Entropy and Mutual Information

Relative Entropy and Mutual Information

Definition 10. The *relative entropy* or *Kullback-Leibler distance* between $p(x)$ and $q(x)$

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \left(\log \frac{p(x)}{q(x)} \right).$$

- **Conventions:** $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$, $p \log \frac{p}{0} = \infty$
- It is not a true distance, since it is not symmetric and does not satisfy the triangle inequality – sometimes called *Kullback-Leibler divergence*.

Definition 11. $(X, Y) \sim p(x, y), p(x), p(y)$ mass functions; the *mutual information* $I(X; Y)$ is the relative entropy between $p(x, y)$ and $p(x)p(y)$:

$$I(X; Y) = D(p(x, y) \parallel p(x)p(y)) \quad (10)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (11)$$

$$= E_{p(x, y)} \left(\log \frac{p(X, Y)}{p(X)p(Y)} \right). \quad (12)$$

Remark. $D(p \parallel q) \neq D(q \parallel p)$, as the next example shown.

Interpretation. $I(X; Y)$ measures the average reduction in uncertainty of X that results from knowing Y .

Example 12. $\mathcal{X} = \{0, 1\}, p(0) = 1 - r, p(1) = r, q(0) = 1 - s, q(1) = s$.

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

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

If $r = s$, then $D(p \parallel q) = D(q \parallel p)$, but for $r = \frac{1}{2}, s = \frac{1}{4}$

$$D(p \parallel q) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} = 0.20752 \text{ bit}$$

$$D(q \parallel p) = \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} = 0.18872 \text{ bit}$$

Example - relative entropy

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

1.4 Relationship between Entropy and Mutual Information

Relationship between Entropy and Mutual Information

Theorem 13 (Mutual information and entropy).

$$I(X; Y) = H(X) - H(Y|X) \quad (13)$$

$$I(X; Y) = H(Y) - H(X|Y) \quad (14)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (15)$$

$$I(X; Y) = I(Y, X) \quad (16)$$

$$I(X, X) = H(X) \quad (17)$$

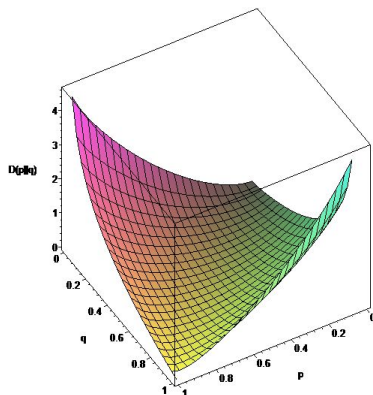


Figure 2: Relative entropy (Kullback-Leibler distance) of two Bernoulli RVs

Proof. (13)

$$\begin{aligned}
 I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)} \\
 &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \underbrace{p(x, y)}_{p(x)} \log p(x) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \\
 &= H(X) - \left(- \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \right) \\
 &= H(X) - H(X|Y)
 \end{aligned}$$

(14) by symmetry

(15) results from (13) and $H(X, Y) = H(Y) - H(X|Y)$; (15) \implies (16)

Finally, $I(X; X) = H(X) - H(X|X) = H(X)$. □

Relationship between entropy and mutual information

Example 14. For the joint distribution of Example 9 the mutual information is

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = 0.375 \text{ bit}$$

The relationship between $H(X)$, $H(Y)$, $H(X, Y)$, $H(X|Y)$, $H(Y|X)$, and $I(X; Y)$ is depicted in Figure 4. Notice that $I(X; Y)$ corresponds to the intersection of the information in X with the information in Y .

Relationship between entropy and mutual information

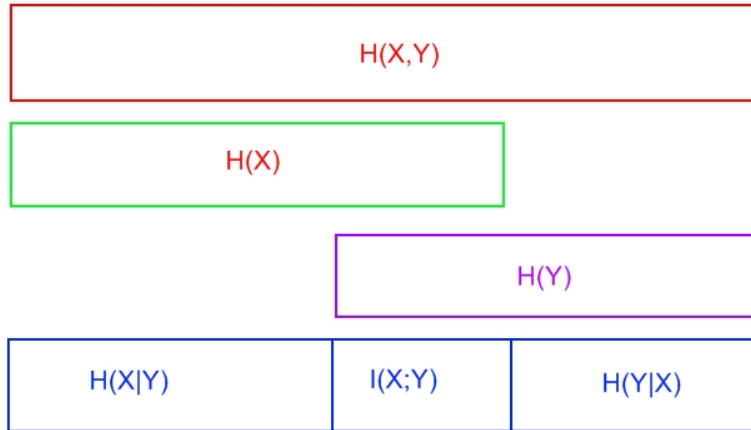


Figure 3: Graphical representation of the relation between entropy and mutual information

Relationship between entropy and mutual information - graphical

1.5 Chain Rules for Entropy, Relative Entropy and Mutual information

Chain rules for entropy, relative entropy and mutual information

Theorem 15 (Chain rule for entropy). $X_1, X_2, \dots, X_n \sim p(x_1, x_2, \dots, x_n)$

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

Proof. Apply repeatedly the two variable expansion rule for entropy

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2 | X_1), \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3 | X_1) \\ &= H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1) \\ &\vdots \\ H(X_1, X_2, \dots, X_n) &= H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_{n-1}, \dots, X_1) \\ &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1). \end{aligned}$$

□

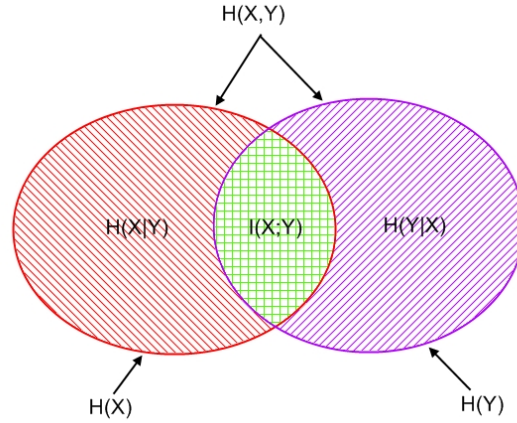


Figure 4: Venn diagram for the relationship between entropy and mutual information

Definition 16. The *conditional mutual information* of random variables X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \quad (18)$$

$$= E_{p(x,y,z)} \log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \quad (19)$$

Theorem 17 (Chain rule for information).

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1). \quad (20)$$

Proof.

$$\begin{aligned} I(X_1, X_2, \dots, X_n; Y) &= H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y) \end{aligned} \quad (21)$$

$$\begin{aligned} &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) - \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1, Y) \\ &= \sum_{i=1}^n I(X_i; Y | X_1, X_2, \dots, X_{i-1}) \end{aligned} \quad (22)$$

□

Definition 18. For joint probability mass functions $p(x, y)$ and $q(x, y)$, the *con-*

ditional relative entropy is

$$D(p(x, y) \parallel q(x, y)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \quad (23)$$

$$= E_{p(x, y)} \log \frac{p(Y|X)}{q(Y|X)}. \quad (24)$$

The notation is not explicit since it omits the mention of the distribution $p(x)$ of the conditioning RV. However it is normally understood from the context.

Theorem 19 (Chain rule for relative entropy).

$$D(p(x, y) \parallel q(x, y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)) \quad (25)$$

Proof.

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

□

2 Inequalities in Information Theory

2.1 Jensen inequality and its consequences

Jensen inequality

Convexity underlies many of the basic properties of information-theoretic quantities such as entropy and mutual information.

Definitions 20. 1. A function $f(x)$ is *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \quad (26)$$

2. f is *strictly convex* if equality holds only for $\lambda = 0$ and $\lambda = 1$.
3. f is *concave* if $-f$ is convex.

- A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.
- Examples of convex functions: x^2 , $|x|$, e^x , $x \log x$ for $x \geq 0$.
- Example of concave functions: $\log x$ and \sqrt{x} for $x \geq 0$.
- If f'' nonnegative (positive) then f is convex (strictly convex)

Theorem 21 (Jensen's inequality). *If f is a convex function and X is a RV*

$$E(f(X)) \geq f(E(X)). \quad (27)$$

If f is strictly convex, equality in (27) implies $X = E(X)$ with probability 1 (i.e. X is a constant).

Proof. for discrete RV induction on number of mass points.

For a two-mass-point distribution, we apply the definition

$$f(p_1x_1 + p_2x_2) \leq p_1f(x_1) + p_2f(x_2)$$

Suppose true for $k - 1$; we set $p'_i = p_i / (1 - p_k)$

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

Extension to continuous distributions using continuity arguments. □

Interpretation of convexity

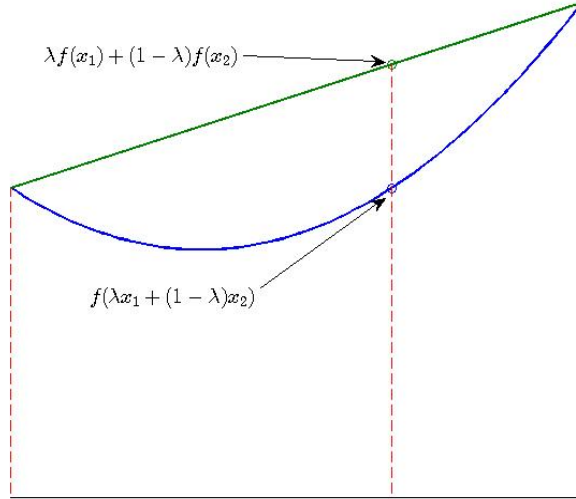
Consequences of Jensen Inequality

- We will use Jensen to prove properties of entropy and relative entropy.

Theorem 22 (Information inequality, Gibbs' inequality). $p(x), q(x), x \in \mathcal{X}$ pmf

$$D(p \parallel q) \geq 0 \quad (28)$$

with equality iff $p(x) = q(x), \forall x \in \mathcal{X}$.



Proof. Let $A := \{x : p(x) > 0\}$

$$\begin{aligned}
 D(p \parallel q) &= \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \left(-\log \frac{q(x)}{p(x)} \right) \\
 &\geq -\log \left(\sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \quad (-\log \text{ is strictly convex}) \\
 &= -\log \sum_{x \in A} q(x) = -\log 1 = 0
 \end{aligned}$$

Equality hold iff $\frac{q(x)}{p(x)} = c, \forall x \in \mathcal{X}$. But, $1 = \sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = c \sum_{x \in \mathcal{X}} p(x) = c$, so $p(x) = q(x), \forall x \in \mathcal{X}$. \square

Since $I(X, Y) = D(p(x, y) \parallel p(x)q(x)) \geq 0$, with equality iff $p(x, y) = p(x)q(x)$ (i.e. X and Y are independent) we obtain

Corollary 23.

$$I(X, Y) \geq 0, \tag{29}$$

with equality iff X and Y are independent.

Corollary 24.

$$I(X; Y|Z) \geq 0, \tag{30}$$

with equality iff X and Y are conditionally independent given Z .

Any random variable over \mathcal{X} has an entropy no greater than $\log |\mathcal{X}|$.

Theorem 25. $H(X) \leq \log |\mathcal{X}|$, with equality iff $X \sim U(\mathcal{X})$.

Proof. $p(x)$ pmf of X , $u(x) = \frac{1}{|\mathcal{X}|}$ pmf of uniform distribution over \mathcal{X} .

$$0 \leq D(p \parallel u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X).$$

□

The next theorem states that conditioning reduces entropy (or information cannot hurt).

Theorem 26.

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

with equality iff X and Y are independent.

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

□

Corollary 27 (Independence bound on entropy).

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

Proof. Chain rule for entropy (Theorem 15)

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \leq \sum_{i=1}^n H(X_i) \quad (\leq \text{ from Th. 26})$$

□

2.2 Log sum inequality and its applications

Log sum inequality and its applications

Theorem 28 (Log sum inequality). a_1, \dots, a_n and b_1, \dots, b_n nonnegative numbers

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \quad (31)$$

with equality iff $\frac{a_i}{b_i} = \text{const.}$

Conventions: $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$ (by continuity)

Proof. Assume w.l.o.g. $a_i > 0, b_i > 0$. Since $f(t) = t \log t$ is convex for $t > 0$, by Jensen ineq.

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right), \quad \alpha_i \geq 0, \sum \alpha_i = 1.$$

Setting $\alpha_i = \frac{b_i}{\sum b_j}$ and $t_i = \frac{a_i}{b_i}$, we obtain

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_j},$$

the desired inequality. \square

Homework. Prove Theorem 22 using log sum inequality.

Using log sum inequality it is easy to prove convexity and concavity results for relative entropy, entropy and mutual information. See [1, Section 2.7].

2.3 Data-processing inequality

Data-processing inequality

Definition 29. 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 depends only on Y and is conditionally independent of X . Specifically, X, Y , and Z form a Markov chain $X \rightarrow Y \rightarrow Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y). \quad (32)$$

Consequences:

- $X \rightarrow Y \rightarrow Z$ iff X and Z are conditionally independent given Y (i.e. $p(x, z|y) = p(x|y)p(z|y)$). Markovity implies conditional independence because

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y). \quad (33)$$

- $X \rightarrow Y \rightarrow Z \implies Z \rightarrow Y \rightarrow X$, sometimes written as $X \longleftrightarrow Y \longleftrightarrow Z$. (reversibility)
- If $Z = f(Y)$, then $X \rightarrow Y \rightarrow Z$

We will prove that no processing of Y , deterministic or random, can increase the information that Y contains about X .

Theorem 30 (Data-processing inequality). *If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$.*

Proof. By chain rule (20), we expand mutual information in two different ways:

$$I(X, Y, Z) = I(X; Z) + I(X; Y|Z) \quad (34)$$

$$= I(X; Y) + I(X; Z|Y). \quad (35)$$

X, Z conditionally independent $\implies I(X; Z|Y) = 0$; since $I(X; Y|Z) \geq 0$ we have

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

We have equality iff $I(X; Y|Z) = 0$, that is $X \rightarrow Z \rightarrow Y$ forms a Markov chain. Similarly, one can prove that $I(Y; Z) \geq I(X; Z)$. \square

Corollary 31. *In particular, if $Z = g(Y)$, we have $I(X; Y) \geq I(X; g(Y))$.*

Proof. $X \rightarrow Y \rightarrow g(Y)$ forms a Markov chain. \square

Functions of the data Y cannot increase the information about X .

Corollary 32. *If $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$.*

Proof. In (34), (35) we have $I(X; Z|Y)$ (by Markovity) and $I(X; Z) \geq 0$. Thus

$$I(X; Y|Z) \leq I(X; Y). \quad (36)$$

\square

If X, Y, Z do not form a Markov chain it is possible that $I(X; Y|Z) > I(X; Y)$. For example, if X and Y are independent fair binary RVs and $Z = X + Y$, then $I(X; Y) = 0$, $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(X|Z)$. But, $H(X|Z) = P(Z = 1)H(X|Z = 1) = \frac{1}{2}$ bit.

2.4 Sufficient statistics

Sufficient statistics

- We apply data-processing inequality in statistics
- $\{f_\theta(x)\}$ family of pmfs, $X \sim f_\theta(x)$, $T(X)$ statistics
- $\theta \rightarrow X \rightarrow T(X)$; data-processing inequality (Theorem 30) implies

$$I(\theta; T(X)) \leq I(\theta; X) \quad (37)$$

with equality when no information is lost.

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

Definition 33. A function $T(X)$ is said to be a *sufficient statistic* relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ (i.e. $\theta \rightarrow X \rightarrow T(X)$ forms a Markov chain).

The definition is equivalent to the condition of equality in data-processing inequality

$$I(\theta; T(X)) = I(\theta; X) \quad (38)$$

for all distributions on θ . Hence sufficient statistics preserve mutual information and conversely.

Examples(sufficient statistics)

1. $X_1, X_2, \dots, X_n, X_i \in \{0, 1\}$, a sequence of i.i.d. Bernoullian variables with parameter $\theta = P(X_i = 1)$. Given n , the number of 1's is a sufficient statistics for θ

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

2. If $X \sim N(\theta, 1)$, that is

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

and X_1, X_2, \dots, X_n is a sample of i.i.d. $N(\theta, 1)$ RVs, then $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is a sufficient statistic.

3. $f_\theta(x)$ pdf for $U(\theta, \theta + 1)$ - a sufficient statistic for θ is

$$T(X_1, \dots, X_n) = (\min\{X_i\}, \max\{X_i\}).$$

Definition 34. A statistic $T(X)$ is a *minimal sufficient statistic* relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistic U ,

$$\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X.$$

Hence, a minimal sufficient statistic maximally compresses the information about θ in the sample.

2.5 Fano's inequality

Fano's inequality

- Suppose we wish to estimate $X \sim p(x)$
- We observe Y related to X by the conditional distribution $p(y|x)$. From Y we calculate $g(Y) = \hat{X}$; \hat{X} is an estimate of X over the alphabet $\hat{\mathcal{X}}$.
- $X \rightarrow Y \rightarrow \hat{X}$ forms a Markov chain
- Define the probability of error

$$P_e = P\{\hat{X} \neq X\}.$$

Theorem 35. For any estimator \hat{X} such that $X \rightarrow Y \rightarrow \hat{X}$, with $P_e = P\{\hat{X} \neq X\}$

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y). \quad (39)$$

This inequality can be weakened to

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y) \quad (40)$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}. \quad (41)$$

Proof. For the first part we define the RV

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X, \\ 0 & \text{if } \hat{X} = X. \end{cases}$$

We expand $H(E, X|\hat{X})$ in two ways using the chain rule

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} \quad (42)$$

$$= \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(E|X, \hat{X})}_{\leq P_e \log |\mathcal{X}|}. \quad (43)$$

□

Proof - continuation. • Since E is a function of X and \hat{X} , $H(E|X, \hat{X}) = 0$.

- $H(E|\hat{X}) \leq H(E) = H(P_e)$ (conditioning reduce entropy)
- Since for $E = 0$, $\hat{X} = X$ and for $E = 1$ entropy is less than the number of possible outcomes

$$\begin{aligned} H(E|X, \hat{X}) &= P(E = 0)H(X|\hat{X}, E = 0) + P(E = 1)H(X|\hat{X}, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \log |\mathcal{X}| \end{aligned} \quad (44)$$

□

Proof - continuation. Combining these results, we obtain

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X})$$

$X \rightarrow Y \rightarrow \hat{X}$ Markov chain $\implies I(X; \hat{X}) \leq I(X; Y) \implies H(X|\hat{X}) \geq H(X|Y)$.

Finally,

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

□

If we set $\widehat{X} = Y$ in Fano's inequality, we obtain

Corollary 36. For any two RVs X and Y , let $p = P(X \neq Y)$. Then

$$H(p) + p \log |\mathcal{X}| \geq H(X|Y). \quad (45)$$

If the estimator $g(Y)$ takes values in \mathcal{X} , we can replace $\log |\mathcal{X}|$ by $\log (|\mathcal{X}| - 1)$.

Corollary 37. Let $P_e = P(X \neq \widehat{X})$, and let $\widehat{X} : \mathcal{Y} \rightarrow \mathcal{X}$; then

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

Proof. Like the proof of Theorem 35, excepting that in (44), the range of possible X outcomes has the cardinal $|\mathcal{X}| - 1$. \square

Remark. Suppose there is no knowledge of Y . Thus, X must be guessed without any information. Let $X \in \{1, 2, \dots, m\}$ and $p_1 \geq p_2 \geq \dots \geq p_m$. Then the best guess of X is $\widehat{X} = 1$ and the resulting probability error is $P_e = 1 - p_1$. Fano's inequality becomes

$$H(P_e) + P_e \log(m - 1) \geq H(X).$$

The pmf

$$(p_1, p_2, \dots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \dots, \frac{P_e}{m-1}\right)$$

achieves this bound with equality: *Fano's inequality is sharp!*

Next results relates probability of error and entropy. Let X and X' be i.i.d. RVs with entropy $H(X)$.

$$P(X = X') = \sum_x p^2(x).$$

Lemma 38. If X and X' are i.i.d. RVs with entropy $H(X)$,

$$P(X = X') \geq 2^{-H(X)}, \quad (46)$$

with equality iff X has uniform distribution.

Proof. Suppose $X \sim p(x)$; Jensen implies

$$\begin{aligned} 2^{E(\log p(X))} &\leq E\left(2^{\log p(X)}\right) \\ 2^{-H(X)} &= 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x). \end{aligned}$$

\square

Corollary 39. Let $X \sim p(x)$, $X' \sim r(x)$, independent RVs over \mathcal{X} . Then

$$P(X = X') \geq 2^{-H(p) - D(p||r)} \quad (47)$$

$$P(X = X') \geq 2^{-H(r) - D(r||p)}. \quad (48)$$

Proof.

$$\begin{aligned} 2^{-H(p)-D(p\|r)} &= 2^{\sum p(x) \log p(x) + \sum p(x) \log \frac{r(x)}{p(x)}} \\ &= 2^{\sum p(x) \log r(x)} \end{aligned}$$

From Jensen and convexity of $f(y) = 2^y$ it follows

$$\begin{aligned} 2^{-H(p)-D(p\|r)} &\leq \sum p(x) 2^{\log r(x)} \\ &= \sum p(x) r(x) = P(X = X'). \end{aligned}$$

□

References

References

- [1] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, 2nd edition, Wiley, 2006.
- [2] David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
- [3] Robert M. Gray, *Entropy and Information Theory*, Springer, 2009