Entropy, Relative Entropy, and Mutual Information

Some basic notions of Information Theory

Radu Trîmbiţaş

UBB

October 2012

Outline

Entropy of a discrete RV

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

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

Definition 1

The entropy of the discrete random variable X

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

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

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

Philosophy - Why log I

- We define the information in event $\{X = x\}$ as $I(X = x) = I(x) = \log \frac{1}{p(x)}$.
- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
- E.g., other possible functions include $I(x) = p(x)^{\frac{1}{n}}$ for some $n > 0$.
- Another example: $I(x) =$ number of prime factors in $\left\lceil \frac{1}{p(x)} \right\rceil$
- But log, as well will see, has a number of attractions.
- For a distribution on n symbols with probabilities $p = (p_1, p_2, \ldots, p_n)$, let $H(p) = H(p1, p2, \ldots, p_n)$ be the entropy of that distribution.
- Consider any information measure, say $\mathcal{H}(p)$ on p, and consider the following three natural and desirable properties.
  1. $H(p)$ takes its largest value when $p_i = 1/n$ for all i.

Philosophy - Why log II

2. If we define the conditional information as

$$\mathcal{H}(Y|X) := \sum_x p(x)\mathcal{H}(Y|X = x).$$

3. For a distribution on $n+1$ symbols, then if the probability of one is zero, we wish for $\mathcal{H}(p_1, p_2, \ldots, p_n, 0) = \mathcal{H}(p_1, p_2, \ldots, p_n)$

Theorem 2 (Khinchin)

If $\mathcal{H}(p_1, p_2, \ldots, p_n)$ satisfies the above 3 properties for all n and for all p such that $p_i \geq 0$, $\forall i$, and $\sum_i p_i = 1$ (i.e., all probability distributions), then

$$\mathcal{H}(p_1, p_2, \ldots, p_n) = -\lambda \sum_i p_i \log p_i,$$

for $\lambda > 0$.

Theorem 3

If a sequence of symmetric function $H_m(p_1, p_2, \ldots, p_m)$ *satisfies*

1. *Normalization* $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$,
2. *Continuity:* $H_2(p, 1-p)$ *is a continuous function on* p
3. *Grouping:* $H_m(p_1, p_2, \ldots, p_m) =$
   $H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_1}{p_1+p_2}\right)$,

then H_m *must be of the form*

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots$$

Entropy - Properties

Lemma 4

$H(X) \geq 0$

Lemma 5

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

Example 6

Let the RV

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

$$H(X) = -p \log p - (1-p) \log(1-p) =: H(p) \tag{3}$$

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

Entropy - Properties II

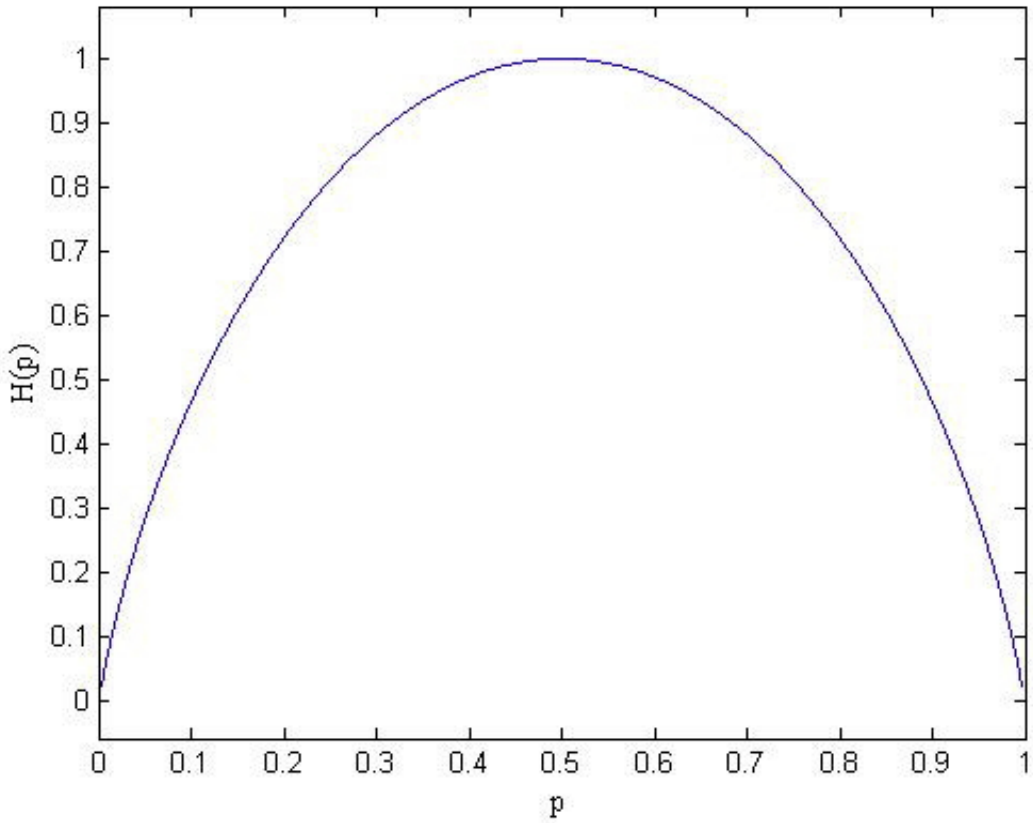


Figure: Graph of $H(p)$

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 $H(X) + 1$.

Joint Entropy and Conditional Entropy I

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

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

- joint distribution of X and Y

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

- (marginal) distribution of X

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

- (marginal) distribution of Y

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

Joint Entropy and Conditional Entropy II

Definition 7

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

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

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

Joint Entropy and Conditional Entropy III

Definition 8

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

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

where

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

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

Joint Entropy and Conditional Entropy IV

- By computation

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

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

$$= -E\left(\log p\left(Y|X\right)\right) \tag{8}$$

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

Joint Entropy and Conditional Entropy V

Theorem 9 (Chain Rule)

$$H(X, Y) = H(X) + H(Y|X). \tag{9}$$

Joint Entropy and Conditional Entropy VI

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. □

Joint Entropy and Conditional Entropy - b I

Corollary 10

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

Example 11

Let (X, Y) have the 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

- marginal distributions X: $\left(\begin{array}{cccc} \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{array}\right)$; Y: $\left(\begin{array}{cccc} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array}\right)$
- $H(X) = \frac{7}{4}$bits, $H(Y) = 2$bits

Joint Entropy and Conditional Entropy - b II

- conditional entropy

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

- $H(Y|X)=\frac{13}{8}$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 $H(X)-H(X|Y)=H(Y)-H(Y|X)$.

Relative Entropy and Mutual Information I

Definition 12

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.
- *Interpretation*: The relative entropy is a measure of the distance between two distributions.
- In statistics, it arises as an expected logarithm of the likelihood ratio.

Relative Entropy and Mutual Information II

- The relative entropy $D(p \parallel q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$.
- If, instead, we used the code for a distribution q, we would need $H(p) + D(p \parallel q)$ bits on the average to describe the random variable.

Relative Entropy and Mutual Information III

Definition 13

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

$$I(X; Y) = D\left(p(x, y) \parallel p(x)p(y)\right) \tag{10}$$

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

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

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.

Relative Entropy and Mutual Information IV

Example 14

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

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

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

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

$$D(p \parallel q) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} = 0.207\,52 \text{ bit}$$

$$D(q \parallel p) = \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} = 0.188\,72 \text{ bit}$$

Example - relative entropy

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

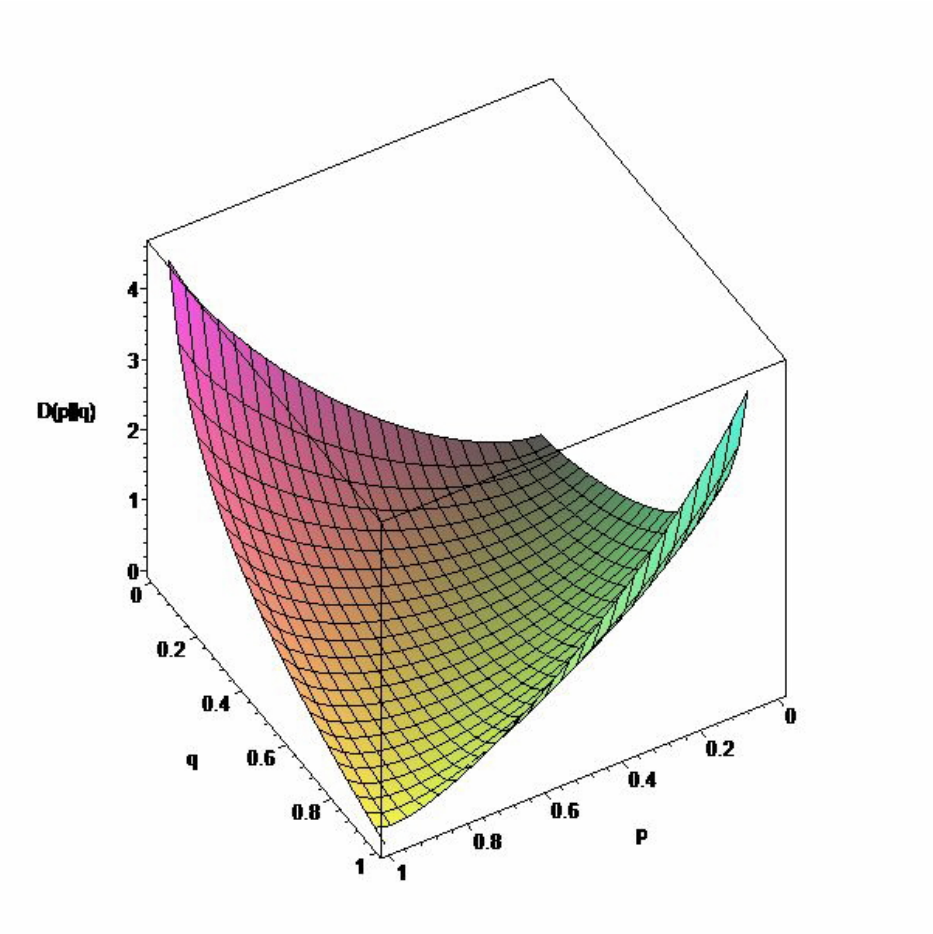


Figure: Relative entropy (Kullback-Leibler distance) of two Bernoulli RVs

Relationship between Entropy and Mutual Information I

Theorem 15 (Mutual information and entropy)

$$I(X;Y) = H(X) - H(Y|X) \tag{13}$$
$$I(X;Y) = H(Y) - H(X|Y) \tag{14}$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{15}$$
$$I(X;Y) = I(Y,X) \tag{16}$$
$$I(X,X) = H(X) \tag{17}$$

Relationship between Entropy and Mutual Information II

Proof.

(13)

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

(14) by symmetry

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

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

Relationship between entropy and mutual information

Example 16

For the joint distribution of Example 11 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

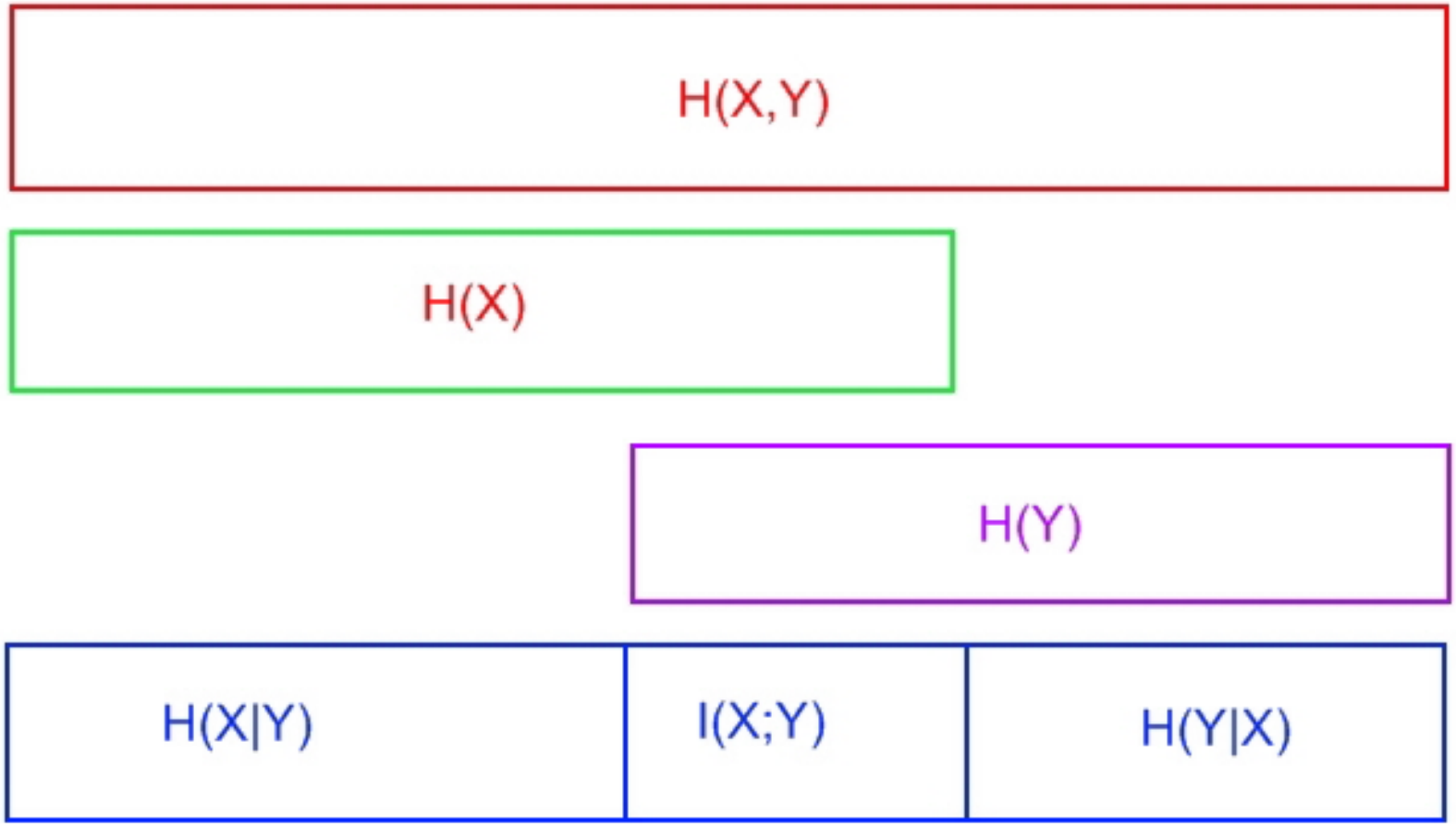


Figure: Graphical representation of the relation between entropy and mutual information

Relationship between entropy and mutual information - graphical

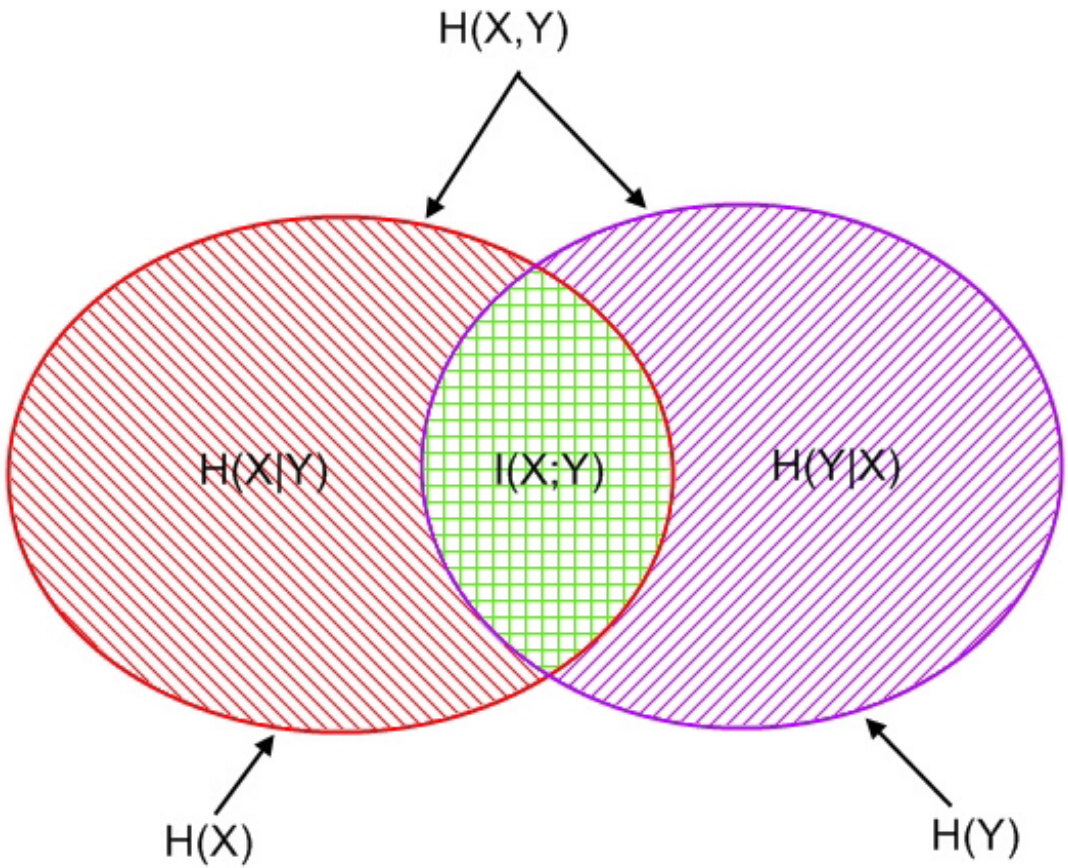


Figure: Venn diagram for the relationship between entropy and mutual information

Chain rules for entropy, relative entropy and mutual information I

Theorem 17 (Chain rule for entropy)

$X_1, X_2, \ldots, X_n \sim p(x_1, x_2, \ldots, x_n)$

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

Chain rules for entropy, relative entropy and mutual information II

Proof.

Apply repeatedly the two variable expansion rule for entropy

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

Chain rules for entropy, relative entropy and mutual information III

Definition 18

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

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{18}$$

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

Theorem 19 (Chain rule for information)

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

Chain rules for entropy, relative entropy and mutual information IV

Proof.

$$
\begin{aligned}
& I(X_1, X_2, \ldots, X_n; Y) \\
&= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y) && (21) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y) \\
&= \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) && (22)
\end{aligned}
$$

Chain rules for entropy, relative entropy and mutual information V

Definition 20

For joint probability mass functions $p(x, y)$ and $q(x, y)$, the conditional relative entropy is

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

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

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.

Chain rules for entropy, relative entropy and mutual information VI

Theorem 21 (Chain rule for relative entropy)

$$D\left(p(x,y) \parallel q(x,y)\right) = D(p(x) \parallel q(x)) + D\left(p(y|x) \parallel q(y|x)\right) \tag{25}$$

Chain rules for entropy, relative entropy and mutual information VII

Proof.

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

Jensen inequality I

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

Definitions 22

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

$$f\left(\lambda x_1 + (1-\lambda) x_2\right) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{26}$$

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

Jensen inequality II

- Example of concave functions: $\log x$ and $\sqrt{x}$ for $x \geq 0$.
- If f'' nonnegative (positive) then f is convex (strictly convex)

Theorem 23 (Jensen's inequality)

If f is a convex function and X is a RV

$$E(f(X)) \geq f(E(X)). \tag{27}$$

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

Jensen inequality III

Proof.

for discrete RV induction on number of mass points.
For a two-mass-point distribution, we apply the definition

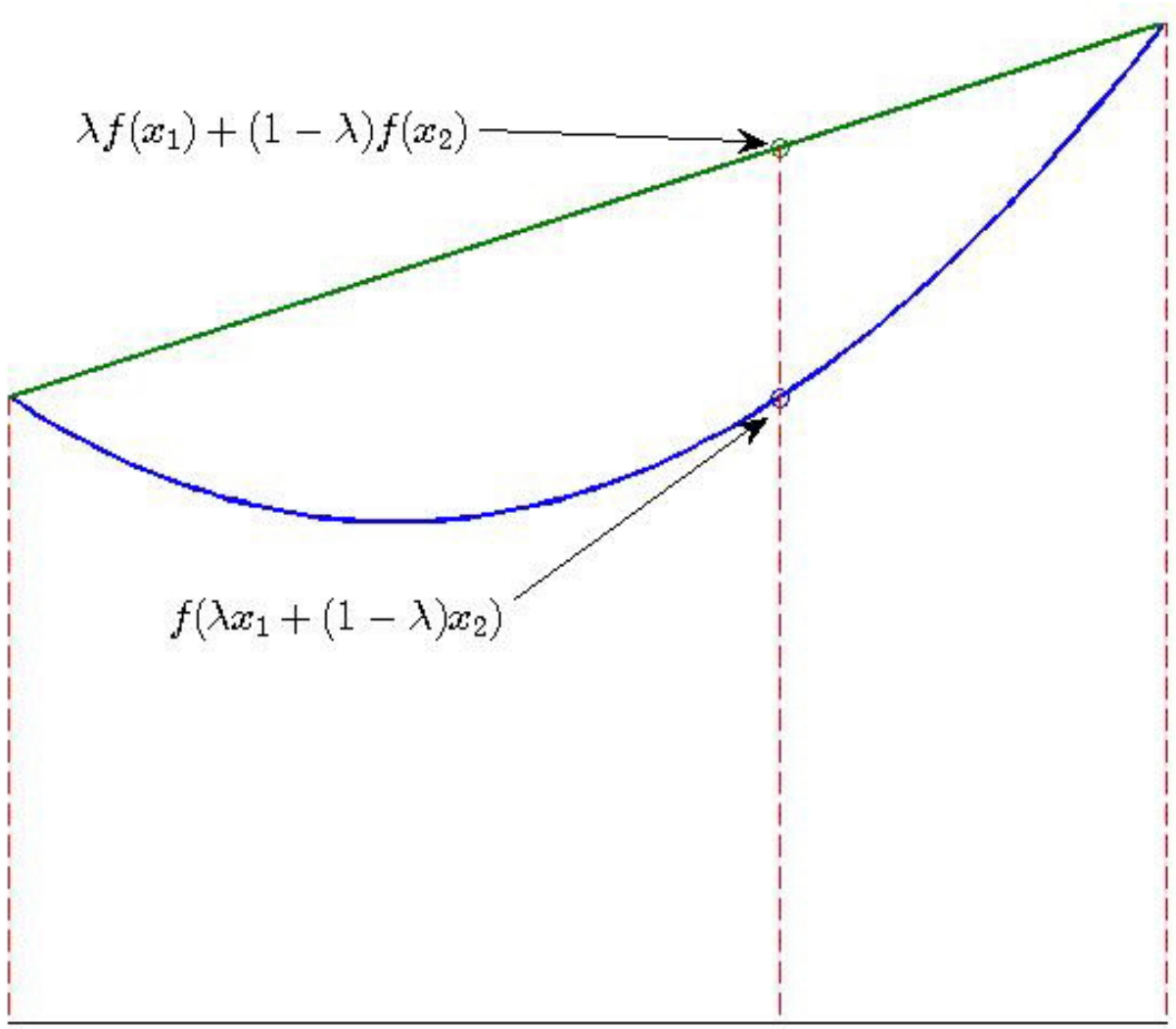
$$f\left(p_1 x_1 + p_2 x_2\right) \leq p_1 f\left(x_1\right) + p_2 f\left(x_2\right)$$

Suppose true for $k-1$; we set $p_i' = p_i / \left(1 - p_k\right)$

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

Extension to continuous distributions using continuity arguments. □

Interpretation of convexity



Consequences of Jensen Inequality I

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

Theorem 24 (Information inequality, Gibbs' inequality)

$p(x)$, $q(x)$, $x \in \mathcal{X}$ *pmf*

$$D(p \parallel q) \geq 0 \tag{28}$$

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

Consequences of Jensen Inequality II

Proof.

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

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

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

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

Consequences of Jensen Inequality III

Corollary 25

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

with equality iff X and Y are independent.

Corollary 26

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

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 27

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

Consequences of Jensen Inequality IV

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 28

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

with equality iff X and Y are independent.

Consequences of Jensen Inequality V

Proof.

$0 \leq I(X, Y) = H(X) - H(X|Y)$. □

Corollary 29 (Independence bound on entropy)

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

Consequences of Jensen Inequality VI

Proof.

Chain rule for entropy (Theorem 17)

$$H(X_1, X_2, \ldots, X) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \leq H(X_i) \quad (\leq \text{ from Th. 28})$$

Log sum inequality and its applications I

Theorem 30 (Log sum inequality)

$a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ nonnegative numbers

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

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

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

Log sum inequality and its applications II

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.

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

Setting $\alpha_i = \dfrac{b_i}{\sum b_j}$ and $t_i = \dfrac{a_i}{b_i}$, 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. □

Homework. Prove Theorem 24 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].

Data-processing inequality I

Definition 31

Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) 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 $X \to Y \to Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y). \tag{32}$$

Consequences:

- $X \to Y \to Z$ 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

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

Data-processing inequality II

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

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

Theorem 32 (Data-processing inequality)

If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Data-processing inequality III

Proof.

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

$$I(X, Y, Z) = I(X; Z) + I(X; Y|Z) \tag{34}$$
$$= I(X; Y) + I(X; Z|Y). \tag{35}$$

X, Z conditionally independent $\Longrightarrow I(X; Z|Y) = 0$; 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 $X \to Z \to Y$ forms a Markov chain. Similarly, one can prove that $I(Y; Z) \geq I(X; Z)$. □

Corollary 33

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

Data-processing inequality IV

Proof.

$X \to Y \to g(Y)$ forms a Markov chain. □

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

Corollary 34

If $X \to Y \to Z$, *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

$$I(X;Y|Z) \leq I(X;Y). \tag{36}$$

□

Data-processing inequality V

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.

Sufficient statistics I

- We apply data-processing inequality in statistics
- $\{f_\theta(x)\}$ family of pmfs, $X \sim f_\theta(x)$, $T(X)$ statistics
- $\theta \to X \to T(X)$; data-processing inequality (Theorem 32) implies

$$I(\theta; T(X)) \leq I(\theta; X) \tag{37}$$

  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.

Sufficient statistics II

Definition 35

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. $\theta \to X \to T(X)$ forms a Markov chain).

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

$$I(\theta; T(X)) = I(\theta; X) \tag{38}$$

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

Examples(sufficient statistics)

Sufficient statistics III

1. $X_1, X_2, \ldots, X_n$, $X_i \in \{0, 1\}$, 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

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

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

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

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

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

$$T(X_1, \ldots, X_n) = \left(\min\{X_i\}, \max\{X_i\}\right).$$

Sufficient statistics IV

Definition 36

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,

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

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

Fano's inequality I

- 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 $g(Y) = \widehat{X}$; $\widehat{X}$ is an estimate of X over the alphabet $\widehat{\mathcal{X}}$.
- $X \to Y \to \widehat{X}$ forms a Markov chain
- Define the probability of error

$$P_e = P\left\{\widehat{X} \neq X\right\}.$$

Fano's inequality II

Theorem 37

For any estimator $\widehat{X}$ *such that* $X \to Y \to \widehat{X}$, *with* $P_e = P\left\{\widehat{X} \neq X\right\}$

$$H(P_e) + P_e \log |\mathcal{X}| \geq H\left(X|\widehat{X}\right) \geq H(X|Y). \tag{39}$$

This inequality can be weakened to

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

or

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

Fano's inequality III

Proof.

For the first part we define the RV

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

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

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

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

. . .

Fano's inequality IV

Proof - continuation.

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

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

. . .

Fano's inequality -continuation I

Proof - continuation.

Combining these results, we obtain

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

$X \to Y \to \widehat{X}$ Markov chain
$\Longrightarrow I(X;\widehat{X}) \leq I(X;Y) \Longrightarrow H(X|\widehat{X}) \geq H(X|Y)$.
Finally,

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

Fano's inequality - consequences I

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

Corollary 38

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

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

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

Corollary 39

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

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

Fano's inequality - consequences II

Proof.

Like the proof of Theorem 37, excepting that in (44), the range of possible X outcomes has the cardinal $|\mathcal{X}| - 1$. □

Remark. Suppose there is no knowledge of Y. Thus, X must be guessed without any information. Let $X \in \{1, 2, \ldots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. 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

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

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

Fano's inequality - consequences III

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 40

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

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

with equality iff X has uniform distribution.

Fano's inequality - consequences IV

Proof.

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

$$2^{E(\log p(X))} \leq E\left(2^{\log P(X)}\right)$$

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x).$$

□

Corollary 41

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

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

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

Fano's inequality - consequences V

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

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