Elements of Information Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1 Entropy, Relative Entropy, and Mutual Information

- Entropy
- Joint Entropy and Conditional Entropy
- Relative Entropy and Mutual Information
- Relationship Between Entropy and Mutual Information
- Chain Rules for Entropy, Relative Entropy and Mutual Information
- Jensen's Inequality and Its Consequences
- Log Sum Inequality and Its Applications
- Data-Processing Inequality
- Sufficient Statistics
- Fano's Inequality

Subsection 1

Entropy

Discrete Random Variables

- We first introduce the concept of entropy, which is a measure of the uncertainty of a random variable.
- Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function

  $$p(x) = \Pr\{X = x\}, \quad x \in \mathcal{X}.$$

  Notation: We denote the probability mass function by $p(x)$ rather than $p_X(x)$, for convenience.

  Thus, $p(x)$ and $p(y)$ refer to two different random variables and are different probability mass functions, $p_X(x)$ and $p_Y(y)$, respectively.

Entropy

Definition

The **entropy** $H(X)$ of a discrete random variable X is defined by

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

- We also write $H(p)$ for the above quantity.
- The log is base 2 and entropy is expressed in **bits**.

  Example: The entropy of a fair coin toss is 1 bit.

  Convention: We will use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $\lim_{x \to 0}(x \log x) = 0$.
- Adding terms of zero probability does not change the entropy.

Other Bases

- If the base of the logarithm is b, we denote the entropy as $H_b(X)$.
- If the base of the logarithm is e, the entropy is measured in **nats**.
- Unless otherwise specified, we will take all logarithms base 2, and hence all the entropies will be measured in bits.

  Remark: Entropy is a functional of the distribution of X.

  It does not depend on the actual values taken by the random variable X, but only on the probabilities.

Entropy and Expectation

- We denote the expectation by E.
- Thus, if $X \sim p(x)$, the expected value of the random variable $g(X)$ is written
$$E_p g(X) = \sum_{x \in \mathcal{X}} g(x)p(x).$$
- If the probability mass function is clear from context, we write $Eg(X)$.
- The entropy of X can be interpreted as the expected value of the random variable $\log \frac{1}{p(X)}$, where X is drawn according to probability mass function $p(x)$,
$$H(X) = E_p \log \frac{1}{p(X)}.$$

Nonnegativity and Change of Base

Lemma

$H(X) \geq 0$.

- $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$.

Lemma

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

- $\log_b p = \log_b a \log_a p$.
- The second property of entropy enables us to change the base of the logarithm in the definition.

  Entropy can be changed from one base to another by multiplying by the appropriate factor.

Example

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

  Then
  $$H(X) = -p\log p - (1-p)\log(1-p) \overset{\text{def}}{=} H(p).$$

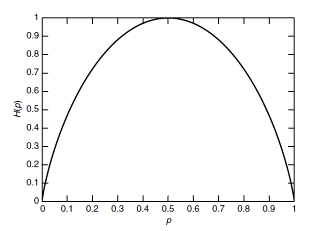
  In particular, $H(X) = 1$ bit when $p = \frac{1}{2}$.

  The graph of the function H(p) is shown in the figure.

  Entropy is a concave function of the distribution.

  It equals 0 when $p = 0$ or 1, i.e., when there is no uncertainty.

  The uncertainty is maximum when $p = \frac{1}{2}$, which also corresponds to the maximum value of the entropy.



Example

- Let $X = \begin{cases} a, & \text{with probability } \frac{1}{2} \\ b, & \text{with probability } \frac{1}{4} \\ c, & \text{with probability } \frac{1}{8} \\ d, & \text{with probability } \frac{1}{8} \end{cases}$.

  The entropy of X is

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

Example (Cont'd)

- Suppose that we wish to determine the value of X with the minimum number of binary questions.
  - An efficient first question is "Is $X = a$?"
    This splits the probability in half.
  - If the answer to the first question is no, the second question can be "Is $X = b$?"
  - The third question can be "Is $X = c$?"

  The resulting expected number of binary questions required is 1.75.

  This turns out to be the minimum expected number of binary questions required to determine the value of X.
- We will show later that the minimum expected number of binary questions required to determine X lies between $H(X)$ and $H(X) + 1$.

Subsection 2

Joint Entropy and Conditional Entropy

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 defined as

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

This can also be expressed as

$$H(X, Y) = -E \log p(X, Y).$$

Conditional Entropy

Definition

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

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

The Chain Rule

Theorem (Chain Rule)

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

- We have

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

Alternatively:

- Write $\log p(X,Y) = \log p(X) + \log p(Y|X)$;
- Take the expectation of both sides to obtain the theorem.

Consequence of the Chain Rule

Corollary

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

- Using the Chain Rule, we obtain

$$\begin{aligned} H(X,Y|Z) &= H(X,Y,Z) - H(Z) \\ &= H(X,Z) - H(Z) + H(X,Y,Z) - H(X,Z) \\ &= H(X|Z) + H(Y|X,Z). \end{aligned}$$

Example

- Let (X, Y) have the joint distribution shown on the right.

$Y \backslash X$	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

- We have:
  - The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.
  - The marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.
  - $H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4}$.
  - $H(Y) = 4(-\frac{1}{4}\log\frac{1}{4}) = 2$.

Example (Cont'd)

- We also have

$$\begin{aligned}
H(X|Y) &= \textstyle\sum_{i=1}^{4} p(Y=i)H(X|Y=i) \\
&= \tfrac{1}{4}H(\tfrac{1}{2},\tfrac{1}{4},\tfrac{1}{8},\tfrac{1}{8}) + \tfrac{1}{4}H(\tfrac{1}{4},\tfrac{1}{2},\tfrac{1}{8},\tfrac{1}{8}) \\
&\qquad + \tfrac{1}{4}H(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}) + \tfrac{1}{4}H(1,0,0,0) \\
&= \tfrac{1}{4}\cdot\tfrac{7}{4} + \tfrac{1}{4}\cdot\tfrac{7}{4} + \tfrac{1}{4}\cdot 2 + \tfrac{1}{4}\cdot 0 \\
&= \tfrac{11}{8} \text{ bits;} \\
H(Y|X) &= \textstyle\sum_{i=1}^{4} p(X=i)H(Y|X=i) \\
&= \tfrac{1}{2}H(\tfrac{1}{4},\tfrac{1}{8},\tfrac{1}{8},\tfrac{1}{2}) + \tfrac{1}{4}H(\tfrac{1}{4},\tfrac{1}{2},\tfrac{1}{4},0) \\
&\qquad + \tfrac{1}{8}H(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{2},0) + \tfrac{1}{8}H(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{2},0) \\
&= \tfrac{1}{2}\cdot\tfrac{7}{4} + \tfrac{1}{4}\cdot\tfrac{3}{2} + \tfrac{1}{8}\cdot\tfrac{3}{2} + \tfrac{1}{8}\cdot\tfrac{3}{2} \\
&= \tfrac{13}{8} \text{ bits.}
\end{aligned}$$

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

Subsection 3

Relative Entropy and Mutual Information

Idea of Relative Entropy

- The entropy of a random variable is a measure of the uncertainty of the random variable.
- The relative entropy is a measure of the distance between two distributions.
- The relative entropy $D(p\|q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.

  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 $H(p) + D(p\|q)$ bits on average to describe the random variable.

The Relative Entropy

Definition

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

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

- In computing relative entropies, we use the following conventions:
  - $0\log\frac{0}{0}=0$;
  - $0\log\frac{0}{q}=0$;
  - $p\log\frac{p}{0}=\infty$.

Remarks on The Relative Entropy

- For a symbol $x \in \mathcal{X}$, such that $p(x) > 0$ and $q(x) = 0$,

$$D(p\|q) = \infty.$$

- We will soon show that:
    - The relative entropy is always nonnegative
    - The relative entropy is zero if and only if $p = q$.
- However, it is not a true distance between distributions since:
    - It is not symmetric;
    - It does not satisfy the triangle inequality.

Mutual Information

- Mutual Information is a measure of the amount of information that one random variable contains about another random variable.
- It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

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 $p(x)p(y)$:

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

Example

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

  Then

$$\begin{aligned} D(p\|q) &= p(0)\log\tfrac{p(0)}{q(0)} + p(1)\log\tfrac{p(1)}{q(1)} \\ &= (1-r)\log\tfrac{1-r}{1-s} + r\log\tfrac{r}{s}; \\ D(q\|p) &= (1-s)\log\tfrac{1-s}{1-r} + s\log\tfrac{s}{r}. \end{aligned}$$

  If $r = s$, then

$$D(p\|q) = D(q\|p) = 0.$$

Example (Cont'd)

- If $r = \frac{1}{2}$, $s = \frac{1}{4}$, we can calculate

$$\begin{array}{rcl}
D(p\|q) & = & \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} \\
 & = & 1 - \frac{1}{2}\log 3 \\
 & = & 0.2075 \text{ bit}; \\
D(q\|p) & = & \frac{3}{4}\log\frac{3/4}{1/2} + \frac{1}{4}\log\frac{1/4}{1/2} \\
 & = & \frac{3}{4}\log 3 - 1 \\
 & = & 0.1887 \text{ bit}.
\end{array}$$

Note that $D(p\|q) \neq D(q\|p)$ in general.

Subsection 4

Relationship Between Entropy and Mutual Information

Mutual Information and Entropy

Theorem (Mutual Information and Entropy)

$$\begin{array}{rcl} I(X;Y) & = & H(X) - H(X|Y) \\ I(X;Y) & = & H(Y) - H(Y|X) \\ I(X;Y) & = & H(X) + H(Y) - H(X,Y) \\ I(X;Y) & = & I(Y;X) \\ I(X;X) & = & H(X). \end{array}$$

- We can rewrite the definition of mutual information $I(X;Y)$ as

$$\begin{array}{rcl} I(X;Y) & = & \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ & = & \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ & = & -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\ & = & -\sum_{x} p(x) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \\ & = & H(X) - H(X|Y). \end{array}$$

Mutual Information and Entropy (Cont'd)

- We showed $I(X;Y) = H(X) - H(X|Y)$.

  Thus, the mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.
- By symmetry, this also follows that $I(X;Y) = H(Y) - H(Y|X)$.

  Thus, X says as much about Y as Y says about X.
- Now recall that $H(X,Y) = H(X) + H(Y|X)$.

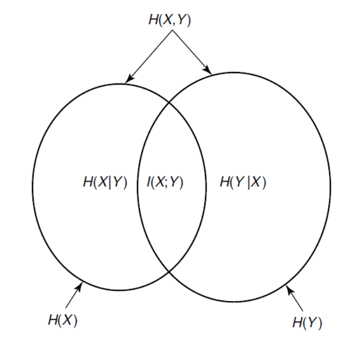
  Thus, we have $I(X;Y) = H(X) + H(Y) - H(X,Y)$.
- Finally, we have $I(X;X) = H(X) - H(X|X) = H(X)$.
- Thus, the mutual information of a random variable with itself is the entropy of the random variable.
- For this reason entropy is sometimes called **self-information**.

Pictorial Representation

- The relationship between

$$H(X), H(Y), H(X,Y),$$
$$H(X|Y), H(Y|X) \text{ and } I(X;Y)$$

is expressed in a Venn diagram.



The mutual information $I(X;Y)$ corresponds to the intersection of the information in X with the information in Y.

Example

- Consider 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

Recall that we calculated $H(X) = \frac{7}{4}$ and $X(X|Y) = \frac{11}{8}$.

Now, it is easy to calculate the mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \frac{14}{8} - \frac{11}{8} \\ &= \frac{3}{8} = 0.375 \text{ bit.} \end{aligned}$$

Subsection 5

Chain Rules for Entropy, Relative Entropy and Mutual Information

Chain Rule for Entropy

Theorem (Chain Rule for Entropy)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

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

- We repeatedly apply the two-variable chain rule.

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

Alternative Proof of the Chain Rule

- We write $p(x_1,\ldots,x_n)=\prod_{i=1}^n p(x_i|x_{i-1},\ldots,x_1)$.

  We then evaluate:

$$\begin{aligned}
&H(X_1,X_2,\ldots,X_n)\\
&= -\sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log p(x_1,x_2,\ldots,x_n)\\
&= -\sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log \prod_{i=1}^n p(x_i|x_{i-1},\ldots,x_1)\\
&= -\sum_{x_1,x_2,\ldots,x_n} \sum_{i=1}^n p(x_1,x_2,\ldots,x_n)\log p(x_i|x_{i-1},\ldots,x_1)\\
&= -\sum_{i=1}^n \sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log p(x_i|x_{i-1},\ldots,x_1)\\
&= -\sum_{i=1}^n \sum_{x_1,x_2,\ldots,x_i} p(x_1,x_2,\ldots,x_i)\log p(x_i|x_{i-1},\ldots,x_1)\\
&= \sum_{i=1}^n H(X_i|X_{i-1},\ldots,X_1).
\end{aligned}$$

Conditional Mutual Information

- We define the conditional mutual information as the reduction in the uncertainty of X due to knowledge of Y when Z is given.

Definition

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

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

Chain Rule for Mutual Information

- Mutual information also satisfies a chain rule.

Theorem (Chain Rule for Information)

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

- We have

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

Conditional Relative Entropy

- We define a conditional version of the relative entropy.

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,

$$\begin{aligned} D(p(y|x)\|q(y|x)) &= \textstyle\sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)}\log\frac{p(Y|X)}{q(Y|X)}. \end{aligned}$$

- The notation for conditional relative entropy is not explicit, since it omits mention of the distribution $p(x)$ of the conditioning random variable, which is normally understood from the context.

The Chain Rule for Relative Entropy

Theorem (Chain Rule for Relative Entropy)

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

- We have

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

Subsection 6

Jensen's Inequality and Its Consequences

Convex and Concave Functions

Definition

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

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

A function f is said to be **strictly convex** if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition

A function 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

- Examples of convex functions include

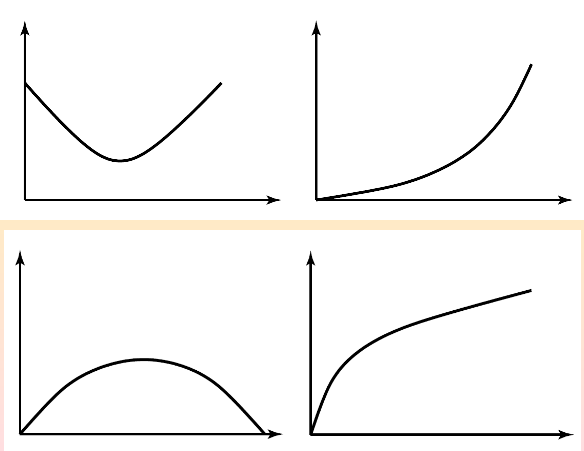
$$x^2, \quad |x|, \quad e^x, \quad x\log x \text{ (for } x \geq 0), \quad \text{and so on.}$$

- Examples of concave functions include

$$\log x \quad \text{and} \quad \sqrt{x}, \text{ for } x \geq 0.$$

Examples

- The figure shows some examples of convex and concave functions.



- Linear functions $ax + b$ are both convex and concave.

Convexity and Second Derivatives

Theorem

If the function f has a second derivative that is nonnegative (positive) over an interval, the function is convex (strictly convex) over that interval.

- We use the Taylor series expansion of the function around x_0:

  $$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2!}(x - x_0)^2,$$

  where x^* lies between x_0 and x. By hypothesis, $f''(x^*) \geq 0$.
  Thus, the last term is nonnegative for all x.
  Letting $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and take $x = x_1$, to obtain

  $$f(x_1) \geq f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)).$$

  Similarly, taking $x = x_2$, we obtain

  $$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)).$$

Convexity and Second Derivatives (Cont'd)

- We got

$$\begin{array}{rcl} f(x_1) & \geq & f(x_0)+f'(x_0)((1-\lambda)(x_1-x_2)), \\ f(x_2) & \geq & f(x_0)+f'(x_0)(\lambda(x_2-x_1)). \end{array}$$

Multiplying the first by λ and the second by $1-\lambda$ and adding, we obtain

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

The proof for strict convexity proceeds along the same lines.

Example: The theorem allows us immediately to verify:

- The strict convexity of x^2, e^x, and $x \log x$, for $x \geq 0$;
- The strict concavity of $\log x$ and $\sqrt{x}$, for $x \geq 0$.

Jensen's Inequality

- Let E denote expectation.
  - $EX = \sum_{x \in \mathcal{X}} p(x)x$ in the discrete case;
  - $EX = \int xf(x)\,dx$ in the continuous case.

Theorem (Jensen's Inequality)

If f is a convex function and X is a random variable,

$$Ef(X) \geq f(EX).$$

Moreover, if f is strictly convex, the equality implies that $X = EX$ with probability 1 (i.e., X is a constant).

- We prove the inequality for discrete distributions using induction on the number of mass points.
- We omit the proof of the second statement.

Jensen's Inequality (Cont'd)

- For a two-mass-point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2).$$

This follows directly from the definition of convex functions.
Suppose the theorem is true for distributions with $k-1$ mass points.
Then writing $p'_i = \frac{p_i}{1-p_k}$, for $i = 1, 2, \ldots, k-1$, we have

$$\begin{array}{rcl}
\sum_{i=1}^{k} p_i f(x_i) & = & p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\
 & \geq & p_k f(x_k) + (1-p_k) f(\sum_{i=1}^{k-1} p'_i x_i) \\
 & & \text{(by the induction hypothesis)} \\
 & \geq & f(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p'_i x_i) \\
 & & \text{(definition of convexity)} \\
 & = & f(\sum_{i=1}^{k} p_i x_i).
\end{array}$$

The proof can be extended to continuous distributions by continuity arguments.

The Information Inequality

Theorem (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 $p(x) = q(x)$, for all x.

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

$$\begin{array}{rcl}
-D(p\|q) & = & -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \\
& = & \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \\
& \leq & \log \sum_{x\in A} p(x) \frac{q(x)}{p(x)} \\
& = & \log \sum_{x\in A} q(x) \\
& \leq & \log \sum_{x\in \mathcal{X}} q(x) \\
& = & \log 1 = 0.
\end{array}$$

The Information Inequality (Cont'd)

- We showed $D(p\|q) \geq 0$.

  Now $\log t$ is a strictly concave function of t.

  It follows that we have equality instead of the first inequality if and only if $\frac{q(x)}{p(x)}$ is constant everywhere, i.e.,

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

  Thus,

  $$\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c.$$

  We have equality in the second inequality only if

  $$\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1.$$

  This implies that $c = 1$.

  Hence, $D(p\|q) = 0$ if and only if $p(x) = q(x)$, for all x.

Consequences

Corollary (Nonnegativity 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.

- We have

$$I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0.$$

Moreover, by the Information Inequality Theorem,

$$\begin{aligned} D(p(x,y) \| p(x)p(y)) = 0 \quad & \text{iff} \quad p(x,y) = p(x)p(y) \\ & \text{iff} \quad X \text{ and } Y \text{ are independent.} \end{aligned}$$

Consequences (Cont'd)

Corollary

We have

$$D(p(y|x)\|q(y|x)) \geq 0,$$

with equality if and only if $p(y|x) = q(y|x)$, for all y and x, such that $p(x) > 0$.

- Also using the Information Inequality Theorem.

Corollary

We have

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

with equality if and only if X and Y are conditionally independent given Z.

- Note that $I(X;Y|Z) = \sum_z p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$. So we may, once more, use the Information Inequality Theorem.

Maximum Entropy

- We now show that the uniform distribution over the range $\mathcal{X}$ is the maximum entropy distribution over this range.
- It follows that any random variable with this range has an entropy no greater than $\log|\mathcal{X}|$.

Theorem

$H(X) \leq \log|\mathcal{X}|$, 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 X.

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

$$\begin{array}{rcl}\log|\mathcal{X}| - H(X) & = & \sum p(x)\log\frac{1}{u(x)} - \sum p(x)\log\frac{1}{p(x)} \\ & = & \sum p(x)\log\frac{p(x)}{u(x)} \\ & = & D(p\|u) \geq 0.\end{array}$$

Conditioning Reduces Entropy

Theorem (Conditioning Reduces Entropy) (Information Cannot Hurt)

We have

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

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

- By the Nonnegativity of Mutual Information,

$$H(X) - H(X|Y) = I(X;Y) \geq 0.$$

- Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X.
  Note: This is true only on the average.
  $H(X|Y=y)$ may be greater than or less than or equal to $H(X)$.
  However, on the average

$$H(X|Y) = \sum_y p(y) H(X|Y=y) \leq H(X).$$

Example

- Let (X, Y) have the joint distribution shown on the right.

  Then

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

$$\begin{array}{rcl} H(X) & = & H(\frac{1}{8}, \frac{7}{8}) = 0.544 \text{ bit;} \\ H(X|Y=1) & = & H(0,1) = 0 \text{ bits;} \\ H(X|Y=2) & = & H(\frac{1}{2}, \frac{1}{2}) = 1 \text{ bit.} \end{array}$$

  We calculate

$$H(X|Y) = \frac{3}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) = 0.25 \text{ bit.}$$

  Thus, the uncertainty in X is:

  - Increased if $Y = 2$ is observed;
  - Decreased if $Y = 1$ is observed.

  But uncertainty decreases on the average.

Independence Bound on Entropy

Theorem (Independence Bound on Entropy)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

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

with equality if and only if the X_i are independent.

- By the chain rule for entropies,

$$\begin{array}{rcl} H(X_1, X_2, \ldots, X_n) & = & \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ & \leq & \sum_{i=1}^{n} H(X_i), \end{array}$$

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

Subsection 7

Log Sum Inequality and Its Applications

Log Sum Inequality

Theorem (Log Sum Inequality)

For nonnegative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

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

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

- We again use the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$, and $0 \log \frac{0}{0} = 0$. These follow from continuity.
- Assume without loss of generality that $a_i > 0$ and $b_i > 0$.

  Note that $f(t) = t \log t \ \Rightarrow \ f'(t) = \log t + \log e \ \Rightarrow \ f''(t) = \frac{1}{t} \log e$.

  So $f(t) = t \log t$ is strictly convex, for all positive t.

Log Sum Inequality (Cont'd)

- Hence, by Jensen's inequality, for $\alpha_i \geq 0$, with $\sum_i \alpha_i = 1$,

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right).$$

Now set $\alpha_i = \frac{b_i}{\sum_{j=1}^n b_j}$ and $t_i = \frac{a_i}{b_i}$.

$$\begin{array}{rcl} \sum \frac{b_i}{\sum b_i} \frac{a_i}{b_i} \log \frac{a_i}{b_i} & \geq & \sum \frac{b_i}{\sum b_i} \frac{a_i}{b_i} \log \sum \frac{b_i}{\sum b_i} \frac{a_i}{b_i} \\ \frac{1}{\sum b_i} \sum a_i \log \frac{a_i}{b_i} & \geq & \frac{\sum a_i}{\sum b_i} \log \frac{\sum a_i}{\sum b_i} \\ \sum a_i \log \frac{a_i}{b_i} & \geq & (\sum a_i) \log \frac{\sum a_i}{\sum b_i}. \end{array}$$

Finally, taking into account that $f(t) = t \log t$ is strictly convex, we get, from Jensen's Inequality, that equality implies $t_i = \frac{a_i}{b_i} = \text{const}$.

Applying the Log Sum Inequality

Claim: $D(p\|q) \geq 0$, with equality if and only if $p(x) = q(x)$.

By the log sum inequality,

$$\begin{array}{rcl} D(p\|q) & = & \sum p(x) \log \frac{p(x)}{q(x)} \\ & \geq & (\sum p(x)) \log \frac{\sum p(x)}{\sum q(x)} \\ & = & 1 \log \frac{1}{1} \\ & = & 0. \end{array}$$

Equality holds if and only if $\frac{p(x)}{q(x)} = c$.

Since both p and q are probability mass functions, $c = 1$.

Hence, $D(p\|q) = 0$ if and only if $p(x) = q(x)$, for all x.

Convexity of Relative Entropy

Theorem (Convexity of Relative Entropy)

$D(p\|q)$ 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, for all $0 \leq \lambda \leq 1$,

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

- We apply the log sum inequality to a term on the left-hand side.

$$\begin{aligned} & D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\ &= \sum (\lambda p_1 + (1-\lambda)p_2) \log \tfrac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \\ &\leq \sum \left[\lambda p_1 \log \tfrac{\lambda p_1}{\lambda p_2} + (1-\lambda)p_2 \log \tfrac{(1-\lambda)p_2}{(1-\lambda)q_2} \right] \\ &= \lambda \sum p_1 \log \tfrac{p_1}{q_1} + (1-\lambda) \sum p_2 \log \tfrac{p_2}{q_2} \\ &= \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2). \end{aligned}$$

Concavity of Entropy

Theorem (Concavity of Entropy)

$H(p)$ is a concave function of p.

- Let u be the uniform distribution on $\mathcal{X}$.

  We know that $H(u) = \log|\mathcal{X}|$ and

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

  By the preceding theorem, relative entropy is convex.

  Therefore, by the displayed equality, H is concave.

Concavity of Entropy (Alternative Proof)

- Let X_1, X_2 be random variables with values in a set A.
  Suppose that:
  - X_1 has distribution p_1;
  - X_2 has distribution p_2.

  Let

  $$\theta = \begin{cases} 1, & \text{with probability } \lambda \\ 2, & \text{with probability } 1-\lambda \end{cases}.$$

  Let $Z = X_\theta$. Then the distribution of Z is $\lambda p_1 + (1-\lambda)p_2$.

  Now conditioning reduces entropy. So $H(Z) \geq H(Z|\theta)$.

  Equivalently,

  $$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2).$$

  So entropy is concave as a function of the distribution.

Concavity, Convexity and Mutual Information

Theorem

Let $(X,Y) \sim p(x,y) = p(x)p(y|x)$. The mutual information $I(X;Y)$ is:

- A concave function of $p(x)$ for fixed $p(y|x)$;
- A convex function of $p(y|x)$ for fixed $p(x)$.

- For the first part, we expand the mutual information

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X=x).$$

By Concavity of Entropy, $H(Y)$ is a concave function of $p(y)$.

If $p(y|x)$ is fixed, then $p(y)$ is a linear function of $p(x)$.

Hence, $H(Y)$ is also a concave function of $p(x)$.

The second term is a linear function of $p(x)$.

Therefore, the difference is a concave function of $p(x)$.

Concavity, Convexity and Mutual Information (Cont'd)

- For the second part, fix $p(x)$.

  Consider two different conditional distributions $p_1(y|x)$ and $p_2(y|x)$.

  The corresponding joint distributions are

  $$p_1(x,y) = p(x)p_1(y|x) \quad \text{and} \quad p_2(x,y) = p(x)p_2(y|x).$$

  Their respective marginals are $p(x)$, $p_1(y)$ and $p(x)$, $p_2(y)$.

  Consider a

  $$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x),$$

  a mixture of $p_1(y|x)$ and $p_2(y|x)$, where $0 \leq \lambda \leq 1$.

  The corresponding joint distribution is also a mixture of the corresponding joint distributions,

  $$p_\lambda(x,y) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y).$$

Concavity, Convexity and Mutual Information (Cont'd)

The distribution of Y is also a mixture,

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda)p_2(y).$$

Let $q_\lambda(x,y) = p(x)p_\lambda(y)$ be the product of the marginal distributions. Then we have

$$q_\lambda(x,y) = \lambda q_1(x,y) + (1-\lambda)q_2(x,y).$$

But the mutual information is the relative entropy between the joint distribution and the product of the marginals.

So we get

$$I(X;Y) = D(p_\lambda(x,y) \| q_\lambda(x,y)).$$

By the Convexity of Relative Entropy, $D(p\|q)$ is convex in (p,q).

So the mutual information is a convex function of the conditional distribution.

Subsection 8

Data-Processing Inequality

Ordered Markov Chains

Definition

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

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

Characterization

- Given, random variables X, Y, Z, $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.

  Suppose, first, that $p(x,y,z) = p(x)p(y|x)p(z|y)$.

  Then we have

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

  Suppose, conversely, that $p(x,z|y) = p(x|y)p(z|y)$.

  Then, we have

  $$\begin{aligned} p(x,y,z) &= p(x,z|y)p(y) = p(x|y)p(z|y)p(y) \\ &= p(x,y)p(z|y) = p(x)p(y|x)p(z|y). \end{aligned}$$

- This characterization of Markov chains 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.

Consequences

- $X \to Y \to Z$ implies that $Z \to Y \to X$.
- Thus, the condition is sometimes written

$$X \leftrightarrow Y \leftrightarrow Z.$$

- If $Z = f(Y)$, then $X \to Y \to Z$.

  The hypothesis implies $p(z|x, y) = p(z|y)$.

  Therefore,

$$p(x, y, z) = p(x)p(y|x)p(z|x, y) = p(x)p(y|x)p(z|y).$$

Data-Processing Inequality

- The next theorem demonstrates that no processing of Y, deterministic or random, can increase the information that Y contains about X.

Theorem (Data-Processing Inequality)

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

- By the chain rule, we can expand mutual information in two different ways:

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

But X and Z are conditionally independent given Y.
So we have $I(X;Z|Y) = 0$.
Since $I(X;Y|Z) \geq 0$, we have $I(X;Y) \geq I(X;Z)$.
Equality holds if and only if $I(X;Y|Z) = 0$.
That is, if and only if $X \to Z \to Y$ forms a Markov chain.
Similarly, one can prove that $I(Y;Z) \geq I(X;Z)$.

Effect of Functions of the Data

Corollary

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

- $X \to Y \to g(Y)$ forms a Markov chain.
- Thus functions of the data Y cannot increase the information about X.

Effect of Downstream Observations

Corollary

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

- Consider again

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

  By Markovity, $I(X;Z|Y) = 0$. Moreover, $I(X;Z) \geq 0$.

  We obtain $I(X;Y|Z) \leq I(X;Y)$.

- Thus, the dependence of X and Y is decreased (or remains unchanged) by the observation of a “downstream” random variable Z.

Effect of Downstream Observations

- Note: It is possible that $I(X;Y|Z) > I(X;Y)$ when X, Y and Z do not form a Markov chain.

  Example: Let X and Y be independent fair binary random variables. Let

  $$Z = X + Y.$$

  Then $I(X;Y) = 0$.

  On the other hand,

  $$\begin{array}{rcl} I(X;Y|Z) & = & H(X|Z) - H(X|Y,Z) \\ & = & H(X|Z) \\ & = & P(Z=1)H(X|Z=1) \\ & = & \frac{1}{2} \text{ bit}. \end{array}$$

Subsection 9

Sufficient Statistics

Idea of Sufficient Statistic

- Suppose that we have a family of probability mass functions $\{f_\theta(x)\}$ indexed by θ.
- 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 $\theta \to X \to T(X)$.
- By the data-processing inequality, we have

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

  for any distribution on θ.
- However, if equality holds, no information is lost.
- A statistic $T(X)$ is called sufficient for θ if it contains all the information in X about θ.

Sufficient Statistic

Definition

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.,

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

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 θ.
- Hence sufficient statistics preserve mutual information and conversely.

Example of Sufficient Statistic I

- Let $X_1, X_2, \ldots, X_n$, $X_i \in \{0,1\}$, be an independent and identically distributed (i.i.d.) sequence of coin tosses of a coin with unknown parameter $\theta = \Pr(X_i = 1)$.

  Given n, the number of 1's,
  $$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i,$$
  is a sufficient statistic for θ.

  In fact, we can show that given T, all sequences having that many 1's are equally likely and independent of the parameter θ. Specifically,
  $$\Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) : \textstyle\sum_{i=1}^{n} X_i = k\} = \begin{cases} \frac{1}{\binom{n}{k}}, & \text{if } \sum x_i = k \\ 0, & \text{otherwise} \end{cases}$$
  Thus, $\theta \to \sum X_i \to (X_1, X_2, \ldots, X_n)$ forms a Markov chain,

  This shows that T is a sufficient statistic for θ.

Example of Sufficient Statistic II

- Suppose X is normally distributed with mean θ and variance 1; i.e.,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = \mathcal{N}(\theta, 1).$$

Let $X_1, X_2, \ldots, X_n$ be drawn independently according to this distribution.

A sufficient statistic for θ is the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

It can be verified that the conditional distribution of $X_1, X_2, \ldots, X_n$, conditioned on $\overline{X}_n$ and n does not depend on θ.

Example of Sufficient Statistic III

- Suppose $f_\theta = \text{Uniform}(\theta, \theta+1)$.

  A sufficient statistic for θ is

  $$\begin{aligned} & T(X_1, X_2, \ldots, X_n) \\ & \quad = (\max\{X_1, X_2, \ldots, X_n\}, \min\{X_1, X_2, \ldots, X_n\}). \end{aligned}$$

  Again, one can show (not very easily) that the distribution of the data is independent of the parameter given the statistic T.

Minimal Sufficient Statistic

Definition

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.

Interpreting this in terms of the data-processing inequality, this implies that

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

- A minimal sufficient statistic maximally compresses the information about θ in the sample.
- Other sufficient statistics may contain additional irrelevant information.

  Example: For a normal distribution with mean θ, the pair of functions giving the mean of all odd samples and the mean of all even samples is a sufficient statistic, but not a minimal sufficient statistic.

Subsection 10

Fano's Inequality

Introduction

- Suppose that we know a random variable Y and we wish to guess the value of a correlated random variable X.
- Fano's inequality relates the probability of error in guessing the random variable X to its conditional entropy $H(X|Y)$.
- We can show that the conditional entropy of a random variable X, given another random variable Y, is zero if and only if X is a function of Y.
- Hence, we can estimate X from Y with zero probability of error if and only if $H(X|Y) = 0$.
- Extending this argument, we expect to be able to estimate X with a low probability of error only if the conditional entropy $H(X|Y)$ is small.

Idea of Fano's Inequality

- Suppose that we wish to estimate a random variable X having distribution $p(x)$.
- We observe a random variable Y which is related to X by the conditional distribution $p(y|x)$.
- From Y, we calculate a function $g(Y) = \widehat{X}$, where $\widehat{X}$ is an estimate of X and takes on values in $\widehat{\mathcal{X}}$.
  - We will not restrict the alphabet $\widehat{\mathcal{X}}$ to be equal to $\mathcal{X}$.
  - We will also allow the function $g(Y)$ to be random.
- We wish to bound the probability that $\widehat{X} \neq X$.
- We observe that $X \to Y \to \widehat{X}$ forms a Markov chain.
- Define the probability of error $P_e = \Pr\{\widehat{X} \neq X\}$.

Fano's Inequality

Theorem (Fano's Inequality)

For any estimator $\widehat{X}$, such that $X \to Y \to \widehat{X}$, with $P_e = \Pr(X \neq \widehat{X})$, we have

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

This inequality can be weakened to

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

Remark: Note that $P_e = 0$ implies that $H(X|Y) = 0$.

- We first ignore the role of Y and prove the first inequality.
- Then we use the data-processing inequality to prove the more traditional form of Fano's inequality, given by the second inequality.

Fano's Inequality (Cont'd)

- Define an error random variable

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

Using the chain rule for entropies, we expand $H(E, X|\widehat{X})$ in two different ways

$$\begin{aligned} H(E, X|\widehat{X}) &= H(X|\widehat{X}) + \underbrace{H(E|X, \widehat{X})}_{=0} \\ &= \underbrace{H(E|\widehat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \widehat{X})}_{\leq P_e \log |\mathcal{X}|}. \end{aligned}$$

Since conditioning reduces entropy, $H(E|\widehat{X}) \leq H(E) = H(P_e)$. Since E is a function of X and $\widehat{X}$, $H(E|X, \widehat{X}) = 0$.

Fano's Inequality (Cont'd)

- The remaining term, $H(X|E,\widehat{X})$, can be bounded by noting that:
  - Given $E=0$, $X=\widehat{X}$;
  - Given $E=1$, we can upper bound the conditional entropy by the log of the number of possible outcomes.

  Therefore, we obtain

$$\begin{aligned} H(X|E,\widehat{X}) &= \Pr(E=0)H(X|\widehat{X},E=0) \\ &\qquad\qquad + \Pr(E=1)H(X|\widehat{X},E=1) \\ &\leq (1-P_e)0+P_e\log|\mathcal{X}|. \end{aligned}$$

  Combining these results, we obtain

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

Fano's Inequality (Conclusion)

- We obtained
$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\widehat{X}).$$
Now $X \to Y \to \widehat{X}$ is a Markov chain.
So, by the Data-Processing Inequality, $I(X;\widehat{X}) \leq I(X;Y)$.
Therefore,
$$H(X|\widehat{X}) = H(X) - I(X;\widehat{X}) \geq H(X) - I(X;Y) = H(X|Y).$$
Thus, we have
$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\widehat{X}) \geq H(X|Y).$$

Consequences

Corollary

For any two random variables X and Y, let $p = \Pr(X \neq Y)$. Then

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

- Let $\widehat{X} = Y$ in Fano's inequality.
- For any two random variables X and Y, if the estimator $g(Y)$ takes values in the set $\mathcal{X}$, we can strengthen the inequality slightly by replacing $\log |\mathcal{X}|$ with $\log (|\mathcal{X}| - 1)$.

Corollary

Let $P_e = \Pr(X \neq \widehat{X})$, and let $\widehat{X} : \mathcal{Y} \to \mathcal{X}$. Then

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

Sharpness of Fano's Inequality

- Suppose that 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 $\widehat{X} = 1$.

  The resulting probability of error is $P_e = 1 - p_1$.

  Fano's inequality becomes

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

  The probability mass function

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

  achieves this bound with equality.

  Thus, Fano's inequality is sharp.

Error and Entropy for Two iid Random Variables

- Let X and X' be two independent identically distributed random variables with entropy $H(X)$. The probability at $X = X'$ is given by $\Pr(X = X') = p^2(x)$.

Lemma

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

$$\Pr(X = X') \geq 2^{-H(X)},$$

with equality if and only if X has a uniform distribution.

- Suppose that $X \sim p(x)$.
  By Jensen's inequality, we have $2^{E \log p(X)} \leq E2^{\log p(X)}$.
  This implies that

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

Error and Entropy for Two Random Variables

Corollary

Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then

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

- We have

$$\begin{array}{rcl} 2^{-H(p)-D(p\|r)} & = & 2^{\sum p(x)\log p(x)+\sum p(x)\log\frac{r(x)}{p(x)}} \\ & = & 2^{\sum p(x)\log r(x)} \\ & \leq & \sum p(x)2^{\log r(x)} \\ & = & \sum p(x)r(x) \\ & = & \Pr(X = X'). \end{array}$$

The inequality follows from Jensen's inequality and the convexity of the function $f(y) = 2^y$.