Chapter 2: Entropy and Mutual Information

Chapter 2 outline

- Definitions
- Entropy
- Joint entropy, conditional entropy
- Relative entropy, mutual information
- Chain rules
- Jensen's inequality
- Log-sum inequality
- Data processing inequality
- Fano's inequality

Definitions

A discrete random variable X takes on values x from the discrete alphabet $\mathcal{X}$.

The probability mass function (pmf) is described by

$$p_X(x) = p(x) = \Pr\{X = x\}, \text{ for } x \in \mathcal{X}.$$

The joint pmf of two random variables X and Y taking on values in alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively is described by

$$p_{X,Y}(x, y) = p(x, y) = \Pr\{X = x, Y = y\}, \text{ for } x, y \in \mathcal{X} \times \mathcal{Y}.$$

If $p_X(X = x) > 0$, the conditional probability that the outcome $Y = y$ given that $X = x$ is defined as

$$p_{Y|X}(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Definitions

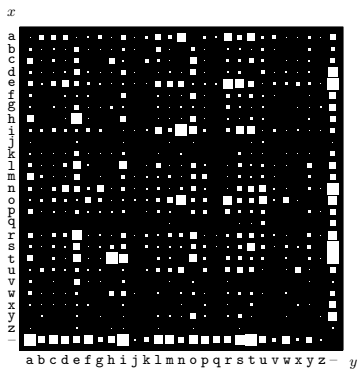


Figure 2.2. The probability distribution over the 27×27 possible bigrams xy in an English language document, *The Frequently Asked Questions Manual for Linux*.

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

Definitions

The events $X = x$ and $Y = y$ are *statistically independent* if $p(x, y) = p(x)p(y)$.

The random variables X and Y defined over the alphabets $\mathcal{X}$ and $\mathcal{Y}$, resp. are *statistically independent* if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$.

The variables $X_1, X_2, \cdots X_N$ are called *independent* if for all $(x_1, x_2, \cdots, x_N) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \mathcal{X}_N$ we have

$$p(x_1, x_2, \cdots x_N) = \prod_{i=1}^{N} p_{X_i}(x_i).$$

They are furthermore called identically distributed if all variables X_i have the same distribution $p_X(x)$.

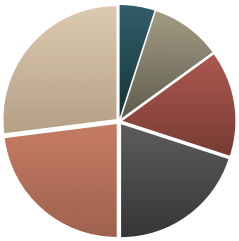
Entropy

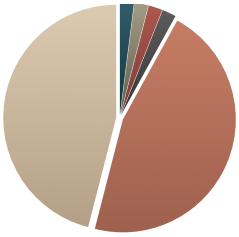
- Intuitive notions?
- 2 ways of defining entropy of a random variable:
  - axiomatic definition (want a measure with certain properties...)
  - just define and then justify definition by showing it arises as answer to a number of natural questions

Definition: The entropy $H(X)$ of a discrete random variable X with pmf $p_X(x)$ is given by

$$H(X) = -\sum_x p_X(x) \log p_X(x) = -E_{p_X(x)}[\log p_X(X)]$$

Order these in terms of entropy





Order these in terms of entropy

University of Illinois at Chicago ECE 534, Fall 2009, Natasha Devroye

Entropy examples 1

- What's the entropy of a uniform discrete random variable taking on K values?

- What's the entropy of a random variable with

$$\mathcal{X} = [\clubsuit, \diamondsuit, \heartsuit, \spadesuit],\ p_X = [1/2; 1/4; 1/8; 1/8]$$

- What's the entropy of a deterministic random variable?

Entropy: example 2

Example 2.12. The entropy of a randomly selected letter in an English document is about 4.11 bits, assuming its probability is as given in table 2.9. We obtain this number by averaging $\log 1/p_i$ (shown in the fourth column) under the probability distribution p_i (shown in the third column).

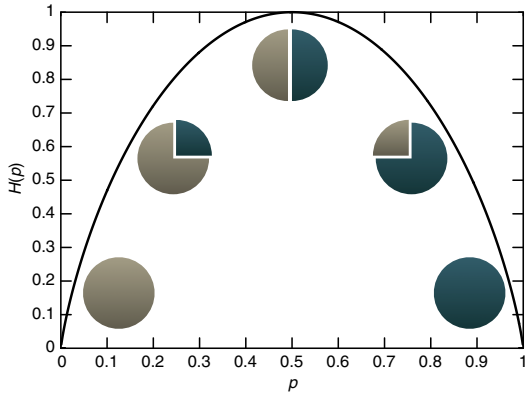
i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
		$\sum_i p_i \log_2 \frac{1}{p_i}$	4.1

Table 2.9. Shannon information contents of the outcomes a–z.

Entropy: example 3

- Bernoulli random variable takes on heads (0) with probability p and tails with probability 1-p. Its entropy is defined as

$$H(p) := -p\log_2(p) - (1-p)\log_2(1-p)$$



Entropy

The entropy $H(X) = -\sum_x p(x)\log p(x)$ has the following properties:

- $H(X) \geq 0$, entropy is always non-negative. $H(X) = 0$ iff X is deterministic $(0\log(0) = 0)$.
- $H(X) \leq log(|\mathcal{X}|)$. $H(X) = log(|\mathcal{X}|)$ iff X has uniform distribution over $\mathfrak{X}$.
- Since $H_b(X) = \log_b(a)H_a(X)$, we don't need to specify the base of the logarithm (bits vs. nat).

Moving on to multiple RVs

Joint entropy and conditional entropy

Definition: Joint entropy of a pair of two discrete random variables X and Y is:

$$\begin{aligned} H(X,Y) &:= -E_{p(x,y)}[\log p(X,Y)] \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \end{aligned}$$

Definition: The conditional entropy of Y given a random variable X (*average* over X)is:

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

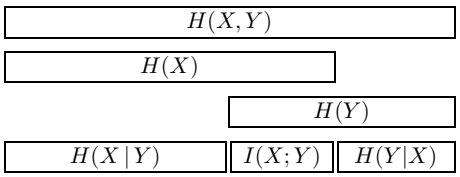
Note: $H(X|Y) \neq H(Y|X)$.

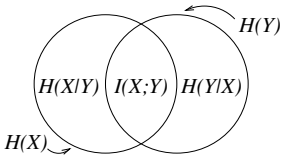
Joint entropy and conditional entropy

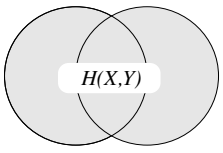
- Natural definitions, since....

Theorem: Chain rule

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







Corollary:

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

Joint/conditional entropy examples

$p(x,y)$	$y=0$	$y=1$
$x=0$	$1/2$	$1/4$
$x=1$	0	$1/4$

H(X,Y)=

H(X|Y)=

H(Y|X)=

H(X)=

H(Y)=

Entropy is central because...

(A) entropy is the measure of **average uncertainty** in the random variable

(B) entropy is the **average number of bits** needed to describe the random variable

(C) entropy is a lower bound on the **average length of the shortest description** of the random variable

(D) entropy is measured in bits?

(E) $H(X) = -\sum_x p(x) \log_2(p(x))$

(F) entropy of a deterministic value is 0

Mutual information

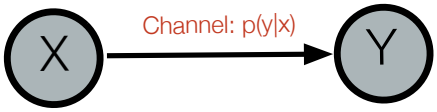
- Entropy H(X) is the uncertainty (`self-information'') of a single random variable
- Conditional entropy H(X|Y) is the entropy of one random variable *conditional upon* knowledge of another.
- The average amount of decrease of the randomness of X by observing Y is the average information that Y gives us about X.

Definition: The mutual information $I(X;Y)$ between the random variables X and Y is given by

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= E_{p(x,y)} \left[\log_2 \frac{p(X,Y)}{p(X)p(Y)} \right] \end{aligned}$$

At the heart of information theory because...

- Information channel capacity:



$$C = \max_{p(x)} I(X;Y)$$

- Operational channel capacity:

```
Highest rate (bits/channel use) that can
communicate at reliably
```

- Channel coding theorem says: information capacity = operational capacity

Mutual information example

$p(x,y)$	$y=0$	$y=1$
$x=0$	$1/2$	$1/4$
$x=1$	0	$1/4$

X or Y	$p(x)$	$p(y)$
0	$3/4$	$1/2$
1	$1/4$	$1/2$

Divergence (relative entropy, K-L distance)

Definition: Relative entropy, divergence or Kullback-Leibler distance between two distributions, P and Q, on the same alphabet, is

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

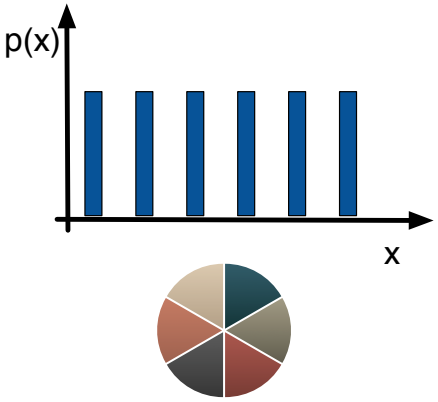
(Note: we use the convention $0 \log \frac{0}{0} = 0$ *and* $0 \log \frac{0}{q} = p \log \frac{p}{0} = \infty$*.)*

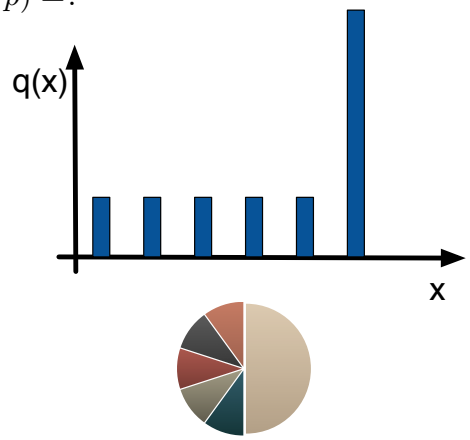
- $D(p \parallel q)$ is in a sense a measure of the "distance" between the two distributions.
- If $P = Q$ then $D(p \parallel q) = 0$.
- *Note* $D(p \parallel q)$ *is not a true distance.*

D(,) = 0.2075 D(,) = 0.1887

K-L divergence example

- $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$
- $P = [1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6]$
- $Q = [1/10 \quad 1/10 \quad 1/10 \quad 1/10 \quad 1/10 \quad 1/2]$
- $D(p \parallel q) =?$ and $D(q \parallel p) =?$





Mutual information as divergence

Definition: The mutual information $I(X;Y)$ between the random variables X and Y is given by

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= E_{p(x,y)} \left[\log_2 \frac{p(X,Y)}{p(X)p(Y)} \right] \end{aligned}$$

- Can we express mutual information in terms of the K-L divergence?

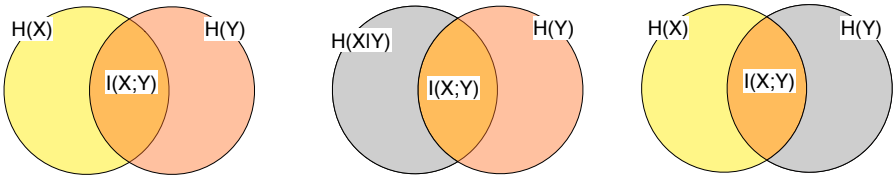
$$I(X;Y) \quad = \quad D(p(x,y) \parallel p(x)p(y))$$

Mutual information and entropy

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$
$$= E_{p(x,y)} \left[\log_2 \frac{p(X,Y)}{p(X)p(Y)} \right]$$

Theorem: Relationship between mutual information and entropy.

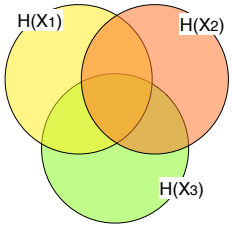
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ I(X;Y) &= H(Y) - H(Y|X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(X;Y) &= I(Y;X) \quad \text{(symmetry)} \\ I(X;X) &= H(X) \quad \text{(“self-information”)} \end{aligned}$$



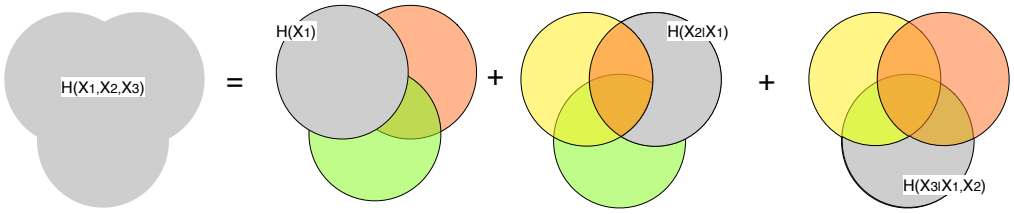
“Two’s company, three’s a crowd”

Chain rule for entropy

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

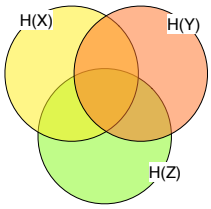


$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1)$$

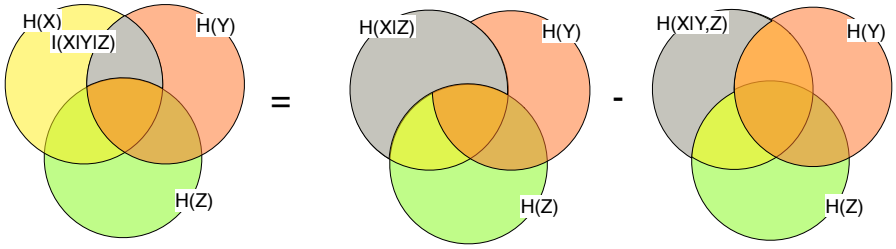


Conditional mutual information

Definition: The conditional mutual information between X and Y given Z is



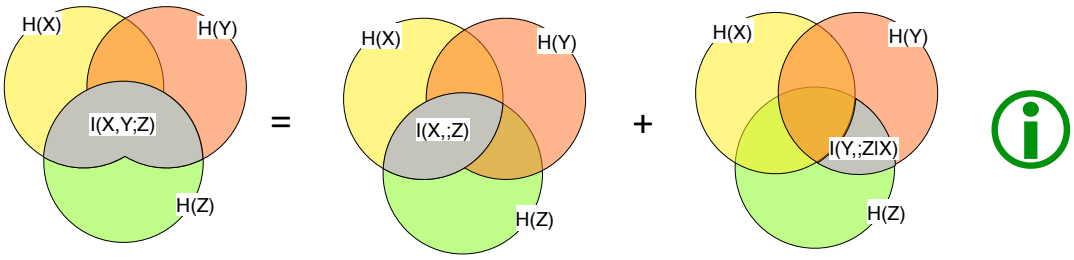
$$\begin{aligned} I(X;Y|Z) &:= H(X|Z) - H(X|Y,Z) \\ &= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$



Chain rule for mutual information

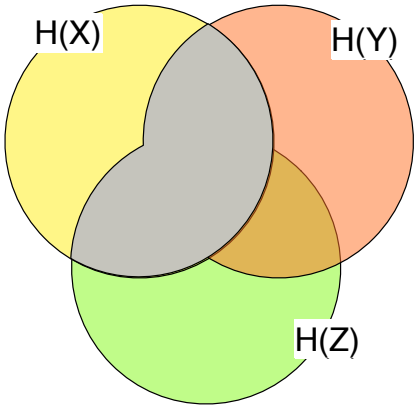
Theorem: (Chain rule for mutual information)

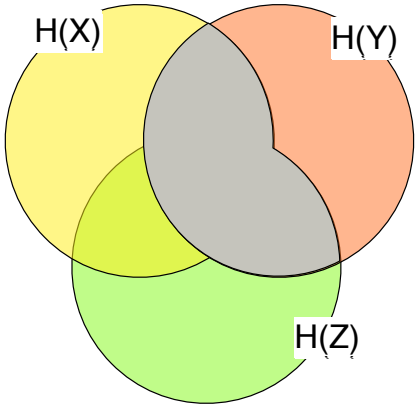
$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1)$$



Chain rule for relative entropy in book pg. 24

What is the grey region?





Another disclaimer....

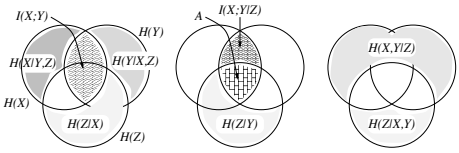


Figure 8.3. A misleading representation of entropies, continued.

that the random outcome (x, y) might correspond to a point in the diagram, and thus confuse entropies with probabilities.

Secondly, the depiction in terms of Venn diagrams encourages one to believe that all the areas correspond to positive quantities. In the special case of two random variables it is indeed true that $H(X|Y)$, $I(X;Y)$ and $H(Y|X)$ are positive quantities. But as soon as we progress to three-variable ensembles, we obtain a diagram with positive-looking areas that may actually correspond to negative quantities. Figure 8.3 correctly shows relationships such as

$$H(X) + H(Z|X) + H(Y|X,Z) = H(X,Y,Z). \quad (8.31)$$

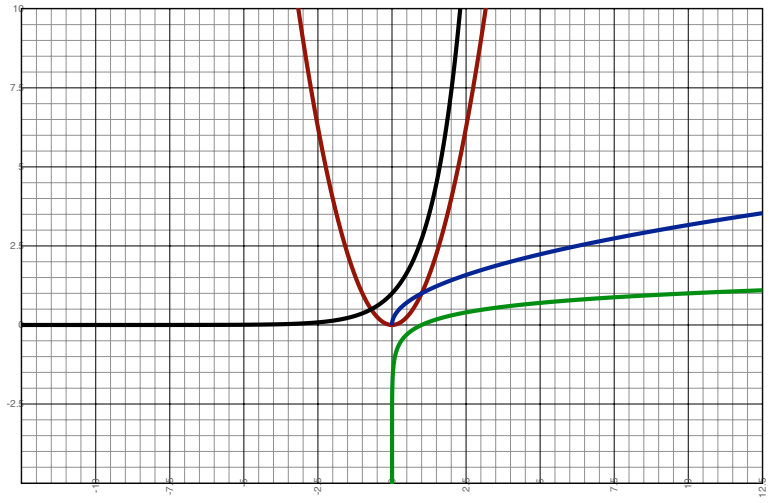
But it gives the misleading impression that the conditional mutual information $I(X;Y|Z)$ is *less than* the mutual information $I(X;Y)$. In fact the area labelled A can correspond to a *negative* quantity. Consider the joint ensemble (X,Y,Z) in which $x \in \{0,1\}$ and $y \in \{0,1\}$ are independent binary variables and $z \in \{0,1\}$ is defined to be $z = x + y \bmod 2$. Then clearly $H(X) = H(Y) = 1$ bit. Also $H(Z) = 1$ bit. And $H(Y|X) = H(Y) = 1$ since the two variables are independent. So the mutual information between X and Y is zero. $I(X;Y) = 0$. However, if z is observed, X and Y become dependent — knowing x, given z, tells you what y is: $y = z - x \bmod 2$. So $I(X;Y|Z) = 1$ bit. Thus the area labelled A must correspond to -1 bits for the figure to give the correct answers.

The above example is not at all a capricious or exceptional illustration. The binary symmetric channel with input X, noise Y, and output Z is a situation in which $I(X;Y) = 0$ (input and noise are independent) but $I(X;Y|Z) > 0$ (once you see the output, the unknown input and the unknown noise are intimately related!).

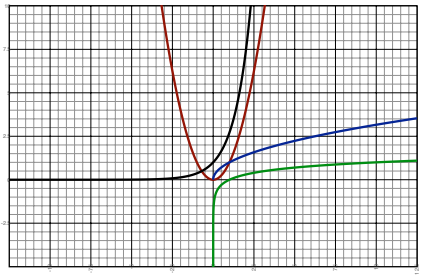
The Venn diagram representation is therefore valid only if one is aware that positive areas may represent negative quantities. With this proviso kept in mind, the interpretation of entropies in terms of sets can be helpful (Yeung, 1991).

[Mackay's textbook]

Convex and concave functions



Convex and concave functions



- A *convex function* f on an interval $[a, b]$ is one for which every chord lies (on or) above the function on that interval.

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v), \quad \forall u, v \in [a, b],\ 0 < \lambda < 1$$

- A function f is *concave* if $-f$ is convex.

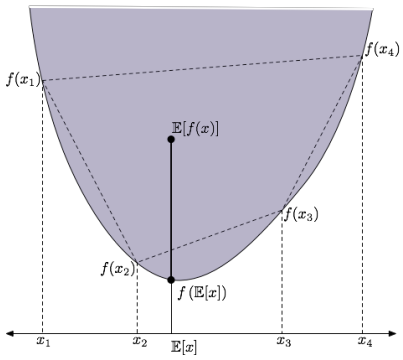
Theorem: If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.

Jensen's inequality

Theorem: (Jensen's inequality) If f is convex, then

$$E[f(X)] \geq f(E[X]).$$

If f is strictly convex, the equality implies $X = E[X]$ with probability 1.



Jensen's inequality consequences

- *Theorem: (Information inequality)* $D(p \parallel q) \geq 0$, with equality iff $p = q$.
- *Corollary: (Nonnegativity of mutual information)* $I(X;Y) \geq 0$ with equality iff X and Y are independent.
- *Theorem: (Conditioning reduces entropy)* $H(X|Y) \leq H(X)$ with equality iff X and Y are independent.
- *Theorem:* $H(X) \leq \log |\mathcal{X}|$ with equality iff X has a uniform distribution over $\mathcal{X}$.
- *Theorem: (Independence bound on entropy)* $H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$with equality iff X_i are independent.

Log-sum inequality

Theorem: (Log sum inequality) For nonnegative $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_n$,

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

with equality iff $a_i/b_i = const$.

Convention: $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$.

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

Log-sum inequality consequences

- *Theorem: (Convexity of relative entropy)* $D(p \parallel q)$ is convex in the pair (p, q), so that for pmf's (p_1, q_1) and (p_2, q_2), we have for all $0 \leq \lambda \leq 1$:

$$\begin{aligned} & D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \\ \leq \quad & \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2) \end{aligned}$$

- *Theorem: Concavity of entropy* For $X \sim p(x)$, we have that

$$H(p) := H_p(X) \text{is a concave function of } p(x).$$

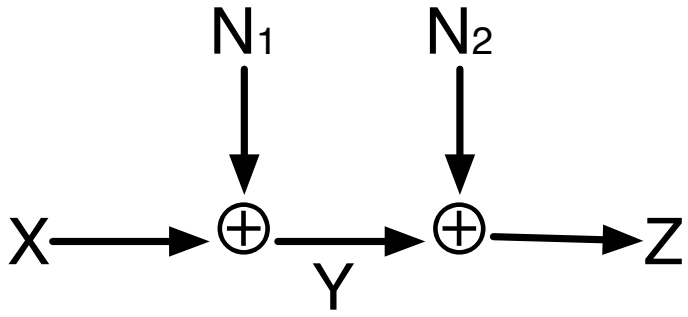
- *Theorem: (Concavity of the mutual information in $p(x)$)* Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$.

- *Theorem: (Convexity of the mutual information in $p(y|x)$)* Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X;Y)$ is a convex function of $p(y|x)$ for fixed $p(x)$.

Markov chains

Definition: X, Y, Z form a Markov chain in that order ($X \to Y \to Z$) iff

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



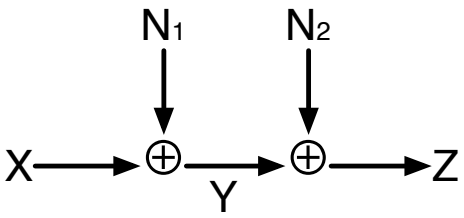
- $X \to Y \to Z$ iff X and Z are conditionally independent given Y
- $X \to Y \to Z \quad \Rightarrow \quad Z \to Y \to X$. Thus, we can write $X \leftrightarrow Y \leftrightarrow Z$.

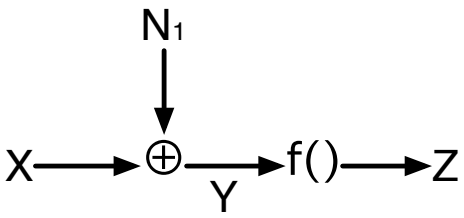
Data-processing inequality

Theorem: (Data-processing inequality) If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$, with equality iff $I(X;Y|Z) = 0$.

Corollary: If $Z = g(Y)$, then $I(X;Y) \geq I(X;g(Y))$.

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





Markov chain questions

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

What if X, Y, Z do not form a Markov chain, can $I(X;Y|Z) \geq I(X;Y)$?

If $X_1 \to X_2 \to X_3 \to X_4 \to X_5 \to X_6$, then Mutual Information increases as you get closer together:

$$I(X_1;X_2) \geq I(X_1;X_4) \geq I(X_1;X_5) \geq I(X_1;X_6).$$

Consequences on sufficient statistics

- Consider a family of probability distributions $\{f_\theta(x)\}$ indexed by θ.
  If $X \sim f(x\,|\,\theta)$ for fixed θ and $T(X)$ is any statistic (i.e., function of the sample X), then we have

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

- The data processing inequality in turn implies

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

  for any distribution on θ.

- Is it possible to choose a statistic that preserves all of the information in X about θ?

Consequences on sufficient statistics

- Consider a family of probability distributions $\{f_\theta(x)\}$ indexed by θ. If $X \sim f(x\,|\,\theta)$ for fixed θ and $T(X)$ is any statistic (i.e., function of the sample X), then we have

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

- The data processing inequality in turn implies

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

for any distribution on θ.

- Is it possible to choose a statistic that preserves all of the information in X about θ?

Definition: Sufficient Statistic A function $T(X)$ is said to be a *sufficient statistic* relative to the family $\{f_\theta(x)\}$ if the conditional distribution of X, given $T(X) = t$, is independent of θ for any distribution on θ (*Fisher-Neyman*):

$$f_\theta(x) = f(x\,|\,t)f_\theta(t) \quad \Rightarrow \quad \theta \to T(X) \to X \quad \Rightarrow \quad I(\theta; T(X)) \geq I(\theta; X)$$

Hence, $I(\theta; X) = I(\theta; T(X))$ for a sufficient statistic.

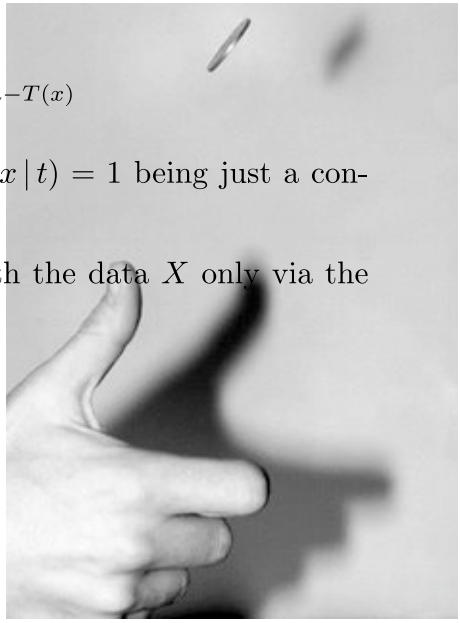
Example of a sufficient statistic

- If $X_1, \ldots, X_n$ are independent Bernoulli-distributed random variables with expected value p, then the sum $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for p.
- Proof: The joint probability distribution

$$p(x_1, \ldots, x_n) = p^{T(x)}(1-p)^{n-T(x)}$$

which satisfies the factorization criterion, with $f(x\,|\,t) = 1$ being just a constant.

- Note that the unknown parameter p interacts with the data X only via the statistic $T(X)$.



Fano's inequality

Theorem: Fano's inequality
For any estimator $\hat{X} : X \to Y \to \hat{X}$, with $P_e = \Pr\{X \neq \hat{X}\}$, we have

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

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

- Fano's inequality says that the probability of error cannot be too small if $H(X|Y)$ is large i.e., correct estimation only happens when the residual randomness of X is small after the observation of Y.

Fano's inequality consequences

- *Corollary:* Let $p = \Pr\{X \neq Y\}$. Then,

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

- *Corollary:* Let $P_e = \Pr\{X \neq \hat{X}\}$, and constrain $\hat{X} : \mathcal{Y} \to \mathcal{X}$; then

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

- Fano's bound is a loose bound, but sufficient for many cases of interest (P_e is small and $|\mathcal{X}|$ is quite large).
- Suppose no observation Y so that X must simply be guessed, and order $X \in \{1, 2, \ldots, m\}$ such that $p_1 \geq p_2 \geq \cdots \geq p_m$. Then $\hat{X} = 1$ is the optimal estimate of X, with $P_e = 1 - p_1$, and 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.