# COMP538: Introduction to Bayesian Networks

## Lecuture 1: Basics of Multivariate Probability and Information Theory

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008

# Objective and outline

- I assume that the students have some exposure to probability theory.
- In this lecture, I will quickly review basic concepts in multivariate probability and information theory. The emphasis will be on intuitions rather than on mathematics.

- Reading: Zhang & Guo, Chapter 1;
- References: Russell & Norvig, Chapter 14; Cover, T. M. and Thomas, J. A (1991). Elements of Information Theory. John Wiley & Sons.

# Outline

# Sample space

- **Sample space (population)** $\Omega$:
  - Set of possible outcomes of some experiment.
  - Example:
    - Experiment: randomly select a student among all UST postgraduate students.
    - Sample space $\Omega$: the set of all UST postgraduate students.
  - Here we assume it to be finite for simplicity.
  - Elements of the sample spaces are called **samples**.

# Events

- Subsets of sample spaces are **events**.
- Examples:
  - Sample space $\Omega$: the set of all UST postgraduate students.
  - $E_{\text{female}} = \{\text{female students}\}$
    the randomly selected student is a female.
  - $E_{\text{male}} = \{\text{male students}\}$
    the randomly selected student is a male.
  - $E_{\text{MPhil}} = \{\text{MPhil students}\}$
    the randomly selected student is an MPhil student.
  - $E_{\text{PhD}} = \{\text{PhD students}\}$
    the randomly selected student is a PhD student.

# Probability measure

- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0 \ \forall A \subseteq \Omega$
3. **Additivity**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

- Example:
  - Sample space $\Omega$: the set of all UST postgraduate students.
  - Define probability measure: $P(A) = |A|/|\Omega|$.
    - $P(E_{\text{female}})$ = 'fraction of female postgraduate students

# Random Variables

- **Random variable $X$**:
  - Function defined over sample space.
  - Example:
    - Gender of (randomly selected) student,
    - Programme of (randomly selected) student
  - Intuitively, a random variable is an unknown quantity.

- **Domain of a random variable $\Omega_X$**:
  - the set of possible states of $X$.
  - Example:

$$\Omega_{\text{Gender}} = \{\text{f}, \text{m}\}$$

,

$$\Omega_{\text{Programme}} = \{\text{PhD}, \text{MPhil}\}$$

.

# Random Variables and Events

- For any state $x$ of a random variable $X$, let

$$\Omega_{X=x} = \{\omega \in \Omega | X(\omega) = x\}$$

  This is an event!

- Example:
  $\Omega_{\text{Gender=f}} = \{$ female postgraduate students in UST$\} = E_{\text{female}}$.
- Note: we use upper case letters, e.g. $X$, for variables and lower case letters, e.g. $x$, for states of variables.
- Note the difference between $\Omega_X$ and $\Omega_{X=x}$

# Probability mass function (distribution)

- **Probability mass function** of a random variable $X$:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

$$P(X = x) = P(\Omega_{X=x})$$

- Examples:
  - $P(\text{Gender=f}) = P(E_{\text{female}}) = 1/6$ (Assumption)
  - $P(\text{Gender=m}) = P(E_{\text{male}}) = 5/6$.
  - $P(\text{Programme=MPhil}) = P(E_{\text{MPhil}}) = 1/3$ (Assumption
  - $P(\text{Programme=PhD}) = P(E_{\text{PhD}}) = 2/3$.
- In practice, we start with probability mass functions, rather than probability measures over sample space $\Omega$.
- Because of Kolmogorov's third axiom, a probability mass function completely determines a probability measure on $\Omega_X$.
- For **continuous** random variable, one has **probability density function** $p(X)$ (here $p$ in lower case).

# Summary

- Sample space: $\Omega$
- Events: $2^{\Omega}$.
- Probability measure:
  - $P : 2^{\Omega} \rightarrow [0, 1]$
  - Three axioms.
- Random variable: $X : \Omega \rightarrow \Omega_X$
- Probability mass function:
  - $P : \Omega_X - > [0, 1]$
  - $P(X = x) = P(\Omega_{X=x})$.
  - Induce probability measure on $2^{\Omega_X}$. Hence we can talk about $P(X \in \{a, b, c\})$.
- $\Omega$ shared by all random variables, enabling us to talk about relationships among them.

# Outline

# Frequentist interpretation

- **Frequentist interpretation**:
- Probability is long term relative frequency

- Example:
  - $X$ is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach $1/2$ as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:
    - $X_i$: result of the i-th tossing; 1 – H, 0 — T
    - Law of Large Numbers:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{2} \qquad \text{with probability 1}$$

- The frequentist interpretation is meaningful only when experiment can be repeated under the same condition.

# Subjectivist/Bayesian interpretation

- Probabilities are logically consistent degrees of beliefs.
- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.
- Example: "probability that Suez canal is longer than the Panama canal".
    - Doesn't make sense under frequentist interpretation.
    - Subjectivist: degree of belief based on state of knowledge
        - Primary school student: 0.5
        - Me: 0.8
        - Geographer: 1 or 0

# Subjectivist/Bayesian interpretation

- Large literature discusses subjcetivist interpretation (see Shafer and Pearl 1990).
- Use betting arguments to prove that degrees of subjective beliefs must satisfy Kolmogorov's axioms.One argument is called **Dutch book**.
- Example: Horse racing
  - Horses: H1, H2, H3
  - Betting tickets:

| | |
|---|---|
| T1: worth 100 if H1 wins | T12: worth 100 if H1 or H2 wins |
| T2: worth 100 if H2 wins | T13: worth 100 if H1 or H3 wins |
| T3: worth 100 if H3 wins | T23: worth 100 if H2 or H3 wins |
| T0:worth 100 if no horse wins | T123: worth 100 if any horse wins |

  - Degrees of beliefs and fair prices of tickets
    - fair price for buying or selling T1 = P(H1 wins) $\times$ 100 + P(H1 loses) $\times$ 0.
    - fair price for buying or selling T2 = P(H2 wins) $\times$ 100
    - fair price for buying or selling T12 = P(H1 or H2 wins) $\times$ 100 ..., etc

# Subjectivist interpretation

- If a person's degrees of beliefs violates Kolmogorov's axioms, a **Dutch book** can be made so that the person will stand to lose regardless of outcome.
- Example:
  - P(H1 wins) = 0.3, P(H2 wins)=0.4, P(H1 or H2 wins) = 0.5

$$P(H1\ or\ H2) < P(H1) + P(H2)$$

  - Dutch book against the person:
    - buy T12 from the person at 50 (this is fair for him),
    - sell T1 and T2 to the person at 30 and 40 (this is also fair for him).
  - Value before and after the transaction:

| | before (T12) | after (T1 & T2) |
|---|---|---|
| H1 wins | 100 | 100 + 50 - 30 - 40=80 |
| H2 wins | 100 | 100 + 50 - 30 - 40=80 |
| H3 wins | 0 | 50 - 30 - 40 =-20 |

  The person loses 20 in the transaction.
- Exercise: What if the other axioms are violated?

# Subjectivist interpretation

- The subjectivist interpretation was not widely accepted in AI until 1970s (Shafer and Pearl 1990,introduction).
- This is a major reason why probability theory did not play a big role in AI before 1980.
  - Because probability was defined as relative statistical frequency and hence was seen as a technique that was appropriate only when statistical data were available.
  - Not many interesting applications with statistical data at that time. Now, more common.

# Subjectivist interpretation

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs (prior probabilities) when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.

  - As we will learn later, probability has a numerical aspect as well as a structural aspect.
    - We will rely more on the subjectivity interpretation when it comes to building structures than estimating numbers. Our belief on "causality" often plays an important role when building structures.

- The subjectivist interpretation makes concepts such as conditional independence easy to understand.

# Outline

# Joint probability mass function

- **Probability mass function** of a random variable $X$:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

$$P(X = x) = P(\Omega_{X=x}).$$

- Suppose there are $n$ random variables $X_1$, $X_2$, ..., $X_n$.
- A **joint probability mass function**, $P(X_1, X_2, \ldots, X_n)$, over those random variables is:
  - a function defined on the Cartesian product of their state spaces:

$$\prod_{i=1}^{n} \Omega_{X_i} \rightarrow [0, 1]$$

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(\Omega_{X_1=x_1} \cap \Omega_{X_2=x_2} \cap \ldots \cap \Omega_{X_n=x_n}).$$

# Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high ($\geq$4k)},
    - Type: {public, private, others}
  - Joint probability distribution $P(\text{Rent}, \text{Type})$:

| | public | private | others |
|---|---|---|---|
| low | .17 | .01 | .02 |
| medium | .44 | .03 | .01 |
| upper medium | .09 | .07 | .01 |
| high | 0 | 0.14 | 0.1 |

# Joint probability distribution

- The joint distribution $P(X_1, X_2, \ldots, X_n)$ contains information about all aspects of the relations among the $n$ random variables.
- In theory, one can answer any query about relations among the variables based on the joint probability.

# Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

P(Type=private) = P(Type=private, Rent=low)+ P(Type=private, Rent=medium)+ P(Type=private, Rent=upper medium)+ P(Type=private, Rent=high)= .25

| | public | private | others | P(Rent) |
|---|---|---|---|---|
| low | .17 | .01 | .02 | .2 |
| medium | .44 | .03 | .01 | .48 |
| upper medium | .09 | .07 | .01 | .17 |
| high | 0 | 0.14 | 0.1 | .15 |
| P(Type) | .7 | .25 | .05 | |

- Called marginal probability because written on the margins.

# Marginal probability

- Write the equations on the previous slide in a compact form:

$$\text{P(Type)} = \sum_{\text{Rent}} \text{P(Type, Rent)}$$

- The operation is called **marginalization**: Variable "Rent" is marginalized from the joint probability P(Type, Rent).

- Notations for more general cases:
  - $$P(X, Y) = \sum_{U,V} P(X, Y, U, V).$$
  - $\mathbf{Y} \subset \{X_1, X_2, \ldots, X_n\}$, $\mathbf{Z} = \{X_1, X_2, \ldots, X_n\} - \mathbf{Y}$,

$$P(\mathbf{Y}) = \sum_{\mathbf{Z}} P(X_1, X_2, \ldots, X_n)$$

# Marginal probability

- A joint probability gives us a full picture about how random variables are related.
- Marginalization lets us to focus one aspect of the picture.

# Conditional probability

- For events $A$ and $B$:

$$P(A|B) = \frac{P(A,B)}{P(B)} (= \frac{P(A \cap B)}{P(B)})$$

- Meaning:
  - $P(A)$: my probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.
- What is the probability of a randomly selected private apartment having "low" rent?

$$\begin{aligned}&P(Rent{=}low|Type{=}private)\\&= \frac{P(Rent{=}Low,\ Type{=}private)}{P(Type{=}private)} = .01/.25 = .04\end{aligned}$$

In contrast:

$$P(Rent{=}low) = 0.2.$$

# Conditional probability

- P(Rent|Type)

| | public | private | others |
|---|---|---|---|
| low | $.17/.7$ | $.01/.25$ | $.02/.05$ |
| medium | $.44/.7$ | $.03/.25$ | $.01/.05$ |
| upper medium | $.09/.7$ | $.07/.25$ | $.01/.05$ |
| high | $0/.7$ | $0.14/.25$ | $0.1/.05$ |

- Note that

$$\sum_{Rent} P(Rent|Type) = 1.$$

- **Notation**: $P(X|Y,Z)$

| X | Y | Z | $P(X\|Y,Z)$ |
|---|---|---|---|
| T | T | T | 0.3 |
| T | T | F | 0.7 |
| : | : | : | : |
| F | F | F | 0.8 |

# Marginal independence

- Two random variables $X$ and $Y$ are **marginally independent**, written $X \perp Y$, if
  - for any state $x$ of $X$ and any state $y$ of $Y$,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of $Y$ does not give me any information about $X$ and vice versa. $Y$ contains no information about $X$ and vice versa.
- Equivalent definition:

$$P(X{=}x, Y{=}y) = P(X{=}x)P(Y{=}y)$$

- Shorthand for the equations:

$$P(X|Y) = P(X), P(X, Y) = P(X)P(Y).$$

# Marginal independence

- Examples:
  - $X$:result of tossing a fair coin for the first time,
    $Y$: result of second tossing of the same coin.
  - $X$: result of US election, $Y$: your grades in this course.
- Counter example:$X$ – oral presentation grade , $Y$ – project report grade.

# Conditional independence

- Two random variables $X$ and $Y$ are **conditionally independent** given a third variable $Z$,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of $Z$ already, then learning the state of $Y$ does not give me additional information about $X$.*
  - $Y$ might contain some information about $X$.
  - However all the information about $X$ contained in $Y$ are also contained in $Z$.
- Shorthand for the equation:

$$P(X|Y, Z) = P(X|Z)$$

- Equivalent definition:

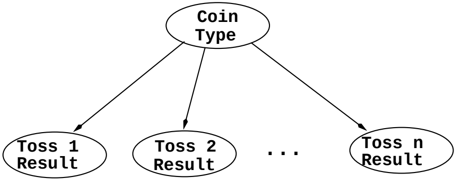
$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

# Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- $X_i$: result of the $i$-th tossing, $Y$: whether the coin is produced by the malfunctioning machine.
- The $X_i$'s are not marginally independent of each other:
  - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.
  - Learning the value of $X_i$ gives me some information about whether the coin is biased, which in term gives me some information about $X_j$.

# Example of Conditional Independence

- However, they are conditionally independent given $Y$:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not, learning the value of $X_i$ does not give me additional information about $X_j$.
- Here is how the variables are related pictorially. We will return to this picture later.



# Equavalent conditions for conditional independence

## Proposition (1.1)

*Variables $X$ and $Y$ are conditionally independent given $Z$ if and only if one of the following conditions is met:*

1. $P(X|Y,Z) = P(X|Z)$ *if* $P(Y,Z) > 0$.
2. $P(X|Y,Z) = f(X,Z)$ *for some functions* $f$.
3. $P(X,Y|Z) = P(X|Z)P(Y|Z)$ *if* $P(Z) > 0$.
4. $P(X,Y|Z) = f(X,Z)g(Y,Z)$ *for some functions* $f$ *and* $g$.
5. $P(X,Y,Z) = P(X|Z)P(Y|Z)P(Z)$ *if* $P(Z) > 0$.
6. $P(X,Y,Z) = P(X,Z)P(Y,Z)/P(Z)$ *if* $P(Z) > 0$.
7. $P(X,Y,Z) = f(X,Z)g(Y,Z)$ **for some functions $f$ and $g$.**

Exercise: Prove the theorem.

# Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: $E = e$
- **Prior probability** $P(H = h)$: belief about a hypothesis before observing evidence.
    - Example: Suppose 10% of people suffer from Hepatitis B. A doctor's prior probability about a new patient suffering from Hepatitis B is 0.1.
- **Posterior probability** $P(H = h|E = e)$:belief about a hypothesis after obtaining the evidence.
    - If the doctor finds that the eyes of the patient are yellow, his belief about patient suffering from Hepatitis B would be $> 0.1$.

# Prior, posterior, and likelihood

- **Likelihood** $L(H = h|E = e)$ of hypothesis $H = h$ given evidence $E = e$
  - Conditional probability of evidence given hypothesis:

$$L(H = h|E = e) = P(E = e|H = h)$$

- Example:
  - Evidence: $E = y$ (Eye-color=yellow);
  - Hypothesis 1: $HB = 1$ (patient has Hepatitis B);
  - Hypothesis 2: $HB = 0$ (patient does not have Hepatitis B);
  - Which hypothesis is more likely given the evidence?
  - Because

$$P(E = y|HB = 1) > P(E = y|HB = 0),$$

    $HB = 1$ is more likely given $E = y$.
  - In general, $P(E = e|H = h)$ measures the likelihood of hypothesis $H = h$.
  - Hence called the likelihood of $H = h$.

# Bayes' Theorem

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(H=h|E=e) = \frac{P(H=h)P(E=e|H=h)}{P(E=e)} \propto P(H=h)P(E=e|H=h)$$

where $P(E=e)$ is normalization constant to ensure $\sum_{h \in \Omega_H} P(H=h|E=e) = 1$.

That is: posterior$(H=h) \propto$ prior$(H=h) \times$ likelihood$(H=h)$

- Example:

$$P(\text{disease}|\text{symptoms}) = \frac{P(\text{disease})P(\text{symptoms}|\text{disease})}{P(\text{symptoms})}$$

  - $P(\text{symptom})$ and $P(\text{symptom}|\text{disease})$ from understanding of disease,
  - $P(\text{disease}|\text{symptoms})$ needed in clinical diagnosis.

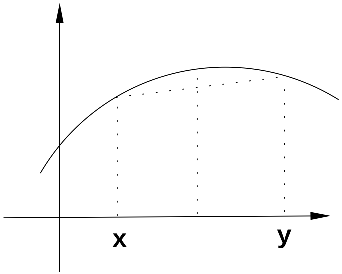
# Outline

# Basics of Information Theory

Review of basics of Information Theory

- Necessary when discussing the use of BN in data analysis,
- Another perspective on conditional independence.

# Concave functions




- A function $f$ is **concave** on interval I if for any $x, y \in I$,

$$\frac{f(x) + f(y)}{2} \leq f(\frac{x+y}{2})$$

Average of function is NO greater than function of average.
It is **strictly concave** if the equality holds only when $x=y$.

# Jensen's Inequality

## Theorem (1.1)

*Suppose function $f$ is concave on interval $I$. Then*

- *For any $p_i \in [0, 1], \sum_{i=1}^{n} p_i = 1$ and $x_i \in I$.*

$$\sum_{i=1}^{n} p_i f(x_i) \leq f(\sum_{i=1}^{n} p_i x_i)$$

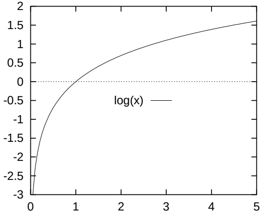
  *Weighted average of function is NO greater than function of weighted average.*

- *If $f$ is strictly CONCAVE, the equality holds iff $p_i \times p_j \neq 0$ implies $x_i = x_j$.*

Exercise: Prove this (using induction).

# Logarithmic function

- The logarithmic function is concave in the interval $(0, \infty)$:




- Hence

$$\sum_{i=1}^{n} p_i log(x_i) \leq log(\sum_{i=1}^{n} p_i x_i) \qquad 0 \leq x_i$$

- In words, exchanging $\sum_i p_i$ with *log* increases a quantity.

# Entropy

- The **entropy** of a random variable $X$:

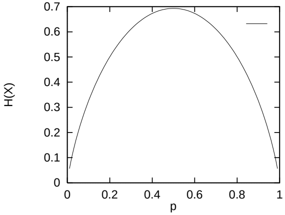
$$H(X) = \sum_{X} P(X) \log \frac{1}{P(X)}$$

with convention that $0 \log(1/0) = 0$.

  - Base of logarithm is 2, unit is bit.
  - Sometimes written as $-E[\log P(x)]$, negation of the expectation of $\log P(X)$.
  - Sometimes, also called the entropy of the distribution.

# Entropy

- $H(X)$ measures uncertainty about $X$:
  - $X$ binary. The chart on the right shows $H(X)$ as a function of $p{=}P(X{=}1)$.
  - The higher $H(X)$ is, the more uncertainty about the value of $X$



# Entropy

Another example:

- $X$ — result of coin tossing
- $Y$ — result of dice throw
- $Z$ — result of randomly pick a card from a deck of 54
- Which one has the highest uncertainty?
- Entropy:

$$\begin{aligned} H(X) &= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1(\log base2) \\ H(Y) &= \frac{1}{6}\log 6 + \ldots + \frac{1}{6}\log 6 = \log 6 \\ H(Z) &= \frac{1}{54}\log 54 + \ldots + \frac{1}{54}\log 54 = \log 54 \end{aligned}$$

Indeed we have:

$$H(X) < H(Y) < H(Z).$$

# Entropy

## Proposition (1.2)

- $H(X) \geq 0$
- $H(X) = 0$ *equality iff* $P(X{=}x) = 1$ *for some* $x \in \Omega_X$*. i.e. iff no uncertainty.*
- $H(X) \leq log(|X|)$ *with equality iff* $P(X{=}x){=}1/|X|$.
  *Uncertainty is the highest in the case of uniform distribution.*

**Proof**: Because *log* is concave, by Jensen's inequality:

$$\begin{aligned} H(X) &= \sum_X P(X) log \frac{1}{P(X)} \\ &\leq log \sum_X P(X) \frac{1}{P(X)} = log|X| \end{aligned}$$

# Conditional entropy

- The **conditional entropy** of $X$ given event $Y{=}y$:
  - Entropy of the conditional distribution $P(X|Y = y)$, i.e.

$$H(X|Y{=}y) \;=\; \sum_{X} P(X|Y{=}y) log \frac{1}{P(X|Y{=}y)}$$

The uncertainty that remains about $X$ when $Y$ is known to be $y$.

- It is possible that $H(X|Y{=}y) > H(X)$
  - Intuitively $Y{=}y$ might contradicts our prior knowledge about $X$ and increase our uncertainty about $X$
  - Exercise: Give example.

# Conditional entropy

- The **conditional entropy** of $X$ given variable $Y$:

$$
\begin{aligned}
H(X|Y) &= \sum_{y \in \Omega_Y} P(Y=y) H(X|Y{=}y) \\
&= \sum_{Y} P(Y) \sum_{X} P(X|Y) log \frac{1}{P(X|Y)} \\
&= \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)} \\
&= -E[log P(X|Y)]
\end{aligned}
$$

The average uncertainty that remains about $X$ when $Y$ is known.

# Joint entropy

- The **joint entropy** of $X$ and $Y$:

$$H(X,Y) = \sum_{X,Y} P(X,Y) log \frac{1}{P(X,Y)}$$

- **Chain rule**:

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

- **Proof**:

$$\begin{aligned}
\sum_{X,Y} P(X,Y) log \frac{1}{P(X,Y)} &= \sum_{X,Y} P(X,Y) log \frac{1}{P(X)P(Y|X)} \\
&= \sum_{X,Y} P(X,Y) log \frac{1}{P(X)} + \sum_{X,Y} P(X,Y) log \frac{1}{P(Y|X} \\
&= \sum_{X} P(X) log \frac{1}{P(X)} + H(Y|X) \\
&= H(X) + H(Y|X)
\end{aligned}$$

# Kullback-Leibler divergence

- **Relative entropy** or **Kullback-Leibler divergence**
  - Measures how much a distribution $Q(X)$ differs from a "true" probability distribution $P(X)$.
  - **K-L divergence** of $Q$ from $P$ is defined as follows:

$$KL(P, Q) = \sum_X P(X) log \frac{P(X)}{Q(X)} = E_P[logP(X)] - E_P[logQ(X)]$$

$0 log \frac{0}{0} = 0$ and $plog \frac{p}{0} = \infty$ if $p \neq 0$

- Not symmetric. So, not a distance measure mathematically.

# Kullback-Leibler divergence

Theorem (1.2)

*(**Gibbs' inequality**)*

$$KL(P,Q) \geq 0$$

*with equality holds iff $P$ is identical to $Q$*

**Proof**:

$$\begin{aligned}
\sum_X P(X) log \frac{P(X)}{Q(X)} &= -\sum_X P(X) log \frac{Q(X)}{P(X)} \\
&\geq -log \sum_X P(X) \frac{Q(X)}{P(X)} \qquad \text{Jensen's inequality} \\
&= -log \sum_X Q(X) = 0.
\end{aligned}$$

KL distance from $P$ to $Q$ is larger than 0 unless $P$ and $Q$ are identical.

# A corollary

## Corollary (1.1)

*Let $f(X)$ be a nonnegative function of variable $X$ such that $\sum_X f(X) > 0$. Let $P^*(X)$ be the probability distribution given by*

$$P^*(X) = \frac{f(X)}{\sum_X f(X)}.$$

*Then for any other probability distribution $P(X)$*

$$\sum_X f(X) log P^*(X) \geq \sum_X f(X) log P(X)$$

*with equality holds iff $P^*$ and $P$ are identical. In other words,*

$$P^* = \arg\sup_P \sum_X f(X) log P(X)$$

# A corollary

**Proof**:

$$KL(P^*, P) = \sum_X P^*(X) log \frac{P^*(X)}{P(X)} \geq 0$$

Hence

$$\sum_X P^*(X) log P^*(X) \geq \sum_X P^*(X) log P(X)$$

$$\sum_X \frac{f(X)}{\sum_X f(X)} log P^*(X) \geq \sum_X \frac{f(X)}{\sum_X f(X)} log P(X)$$

$$\sum_X f(X) log P^*(X) \geq \sum_X f(X) log P(X)$$

Q.E.D

# Mutual information

- The **mutual information** of $X$ and $Y$:

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

- Average reduction in uncertainty about $X$ from learning the value of $Y$, or
- Average amount of information $Y$ conveys about $X$.

# Mutual information and KL Distance

- Note that:

$$
\begin{aligned}
I(X;Y) &= \sum_{X} P(X) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)} \\
&= \sum_{X,Y} P(X,Y) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)} \\
&= \sum_{X,Y} P(X,Y) log \frac{P(X|Y)}{P(X)} \\
&= \sum_{X,Y} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)} \qquad \text{equivalent definition} \\
&= KL(P(X,Y), P(X)P(Y))
\end{aligned}
$$

- Due to equivalent definition:

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

# Property of Mutual information

**Theorem (1.3)**

$$I(X;Y) \geq 0$$

*with equality holds iff* $X \perp Y$.

Interpretation: $X$ and $Y$ are independent iff $X$ contains no information about $Y$ and vice versa.

**Proof**: Follows from previous slide and Theorem 1.2.

# Conditional Entropy Revisited

**Theorem (1.4)**

$H(X|Y) \leq H(X)$ *with equality holds iff* $X \perp Y$

Observation reduces uncertainty in average except for the case of independence.

**Proof**: Follows from Theorem 1.3.

# Mutual information and Entropy

- From definition of mutual information

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

  and the chain rule,

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

  we get

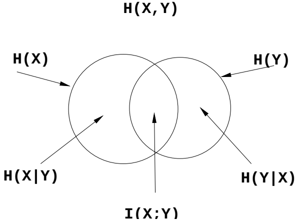
$$H(X) + H(Y) = H(X,Y) + I(X;Y)$$

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

- Consequently
  - $H(X,Y) \leq H(X) + H(Y)$ with equality holds iff $X \perp Y$.

# Mutual information and entropy

Venn Diagram: Relationships among joint entropy, conditional entropy, and mutual information




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

# Conditional Mutual information

- The **conditional mutual information** of $X$ and $Y$ given $Z$:

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

- Average amount of information $Y$ conveys about $X$ given $Z$.

# Conditional mutual information and KL Distance

Note:

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{X,Z} P(X,Z) log \frac{1}{P(X|Z)} - \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Y,Z)} \\
&= \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Z)} - \sum_{X,Y,Z} P(X,Y,Z) log \frac{1}{P(X|Y,Z)} \\
&= \sum_{X,Y,Z} P(X,Y,Z) log \frac{P(X|Y,Z)}{P(X|Z)} \qquad \text{equivalent definition} \\
&= \sum_{Z} P(Z) \sum_{X,Y} P(X,Y|Z) log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \\
&= \sum_{Z} P(Z) KL(P(X,Y|Z), P(X|Z)P(Y|Z)) \geq 0.
\end{aligned}
$$

# Property of conditional mutual information

**Theorem (1.5)**

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

$$H(X|Z) \geq H(X|Y,Z)$$

*with equality hold iff* $X \perp Y|Z$.

Interpretation:

- More observations reduce uncertainty on average except for the case of conditional independence.
- $X$ and $Y$ are independently given $Z$ iff $X$ contain no information about $Y$ given $Z$ and vice versa:

$$X \perp Y|Z \equiv I(X;Y|Z) = 0.$$

Another characterization of conditional independence.