Mathematical Foundations

Berlin Chen

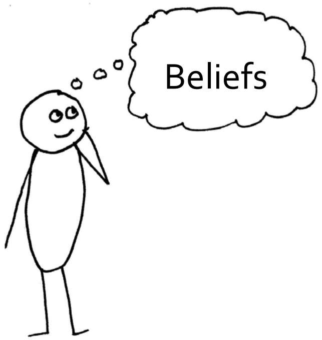
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. C.Manning & H. Schütze, *Foundations of Statistical Natural Language Processing*, Chapter 2
2. D. P. Bertsekas & J. N. Tsitsiklis, *Introduction to Probability*

Role of Probability Theory

- A framework for analyzing phenomena with uncertain outcomes
- Rules for consistent reasoning
- Use for predictions and decisions about the real world



子曰：「由！誨女知之乎！知之為知之，不知為不知，是知也。」

Sources of Images: Microsoft & Google

Experiments, Outcomes and Event

- An **experiment**
  - Produces exactly one out of several possible **outcomes**
  - The set of all possible outcomes is called the **sample space** of the experiment, denoted by
  - A subset of the sample space (a collection of possible outcomes) is called an **event**

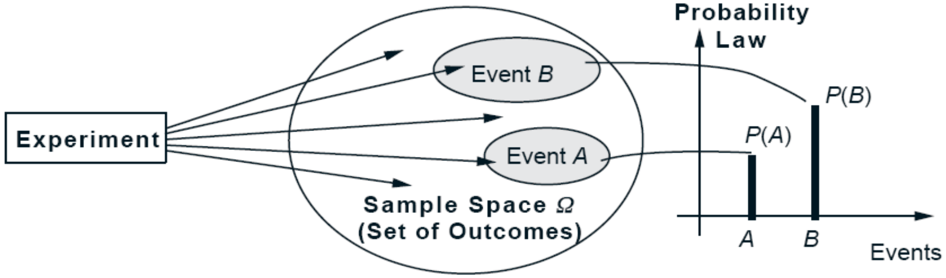
- Examples of the **experiment**
  - A single toss of a coin (finite outcomes)
  - Three tosses of two dice (finite outcomes)
  - An infinite sequences of tosses of a coin (infinite outcomes)
  - Throwing a dart on a square (infinite outcomes), etc.

outcomes:
H, T

events
{H, T}, {H}, {T}, Ø

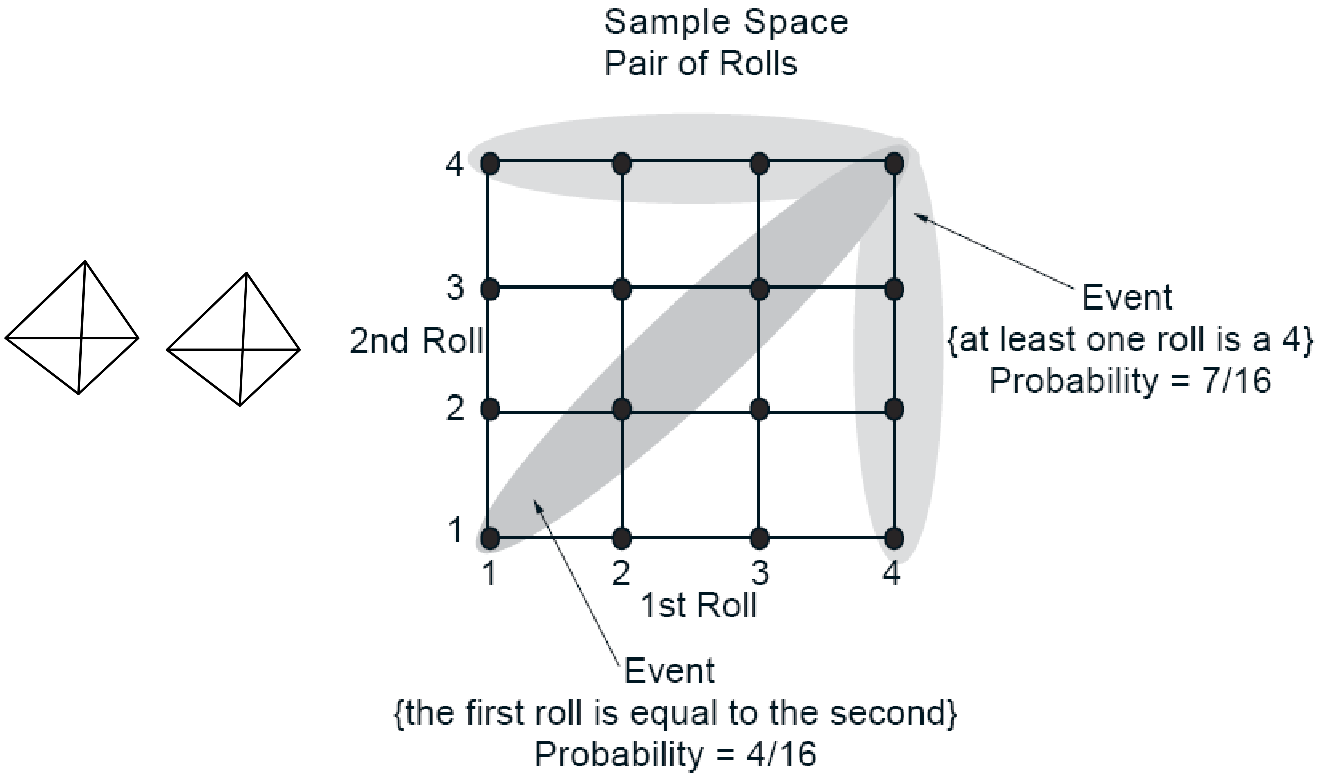
Probabilistic Models

- A probabilistic model is a mathematical description of an uncertainty situation or an experiment
- Elements of a probabilistic model
  - The **sample space**
    - The set of all possible outcomes of an experiment
  - The **probability law**
    - Assign to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective "likelihood" of the elements of



An Example of Sample Space and Probability Law

- The experiment of rolling a pair of 4-sided dice



Sample Space
Pair of Rolls

Event
{at least one roll is a 4}
Probability = 7/16

Event
{the first roll is equal to the second}
Probability = 4/16

Granularity of the Sample Space

Ω

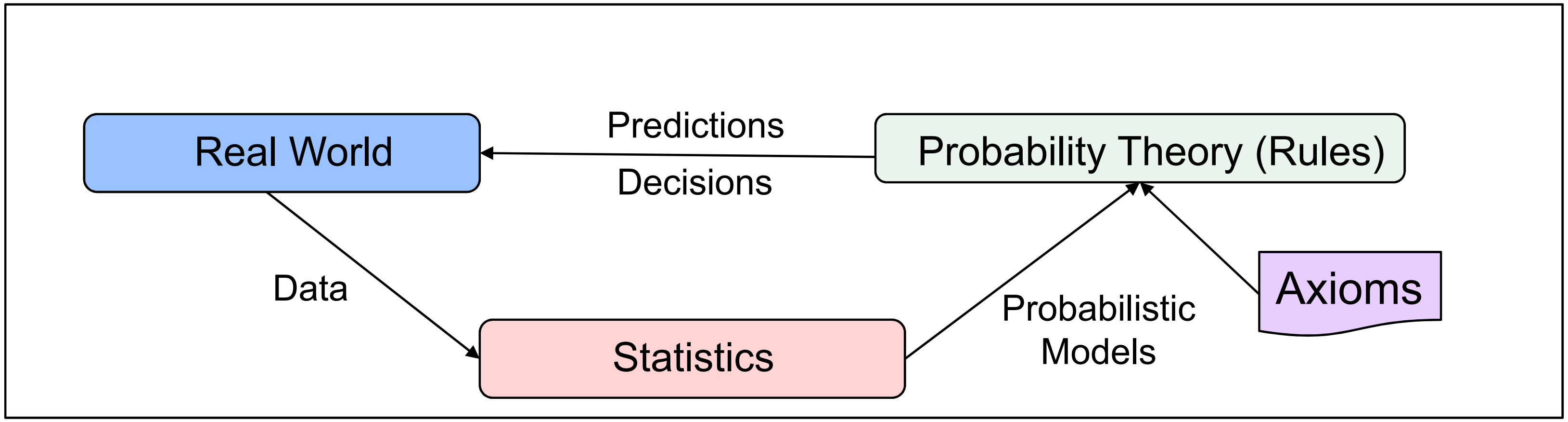
● Head ● Tail

Ω

- ● Head and Rains
- ● Head and No Rains
- ● Tail and Rains
- ● Tail and No Rains

What is the notion of “**Abstraction**”?

Probability and Statistics



Three Probability Axioms

- **Nonnegativity**
  - $\mathbf{P}(A) \geq 0$, for every event A
- **Additivity**
  - If A and B are two disjoint events, then the probability of their union satisfies

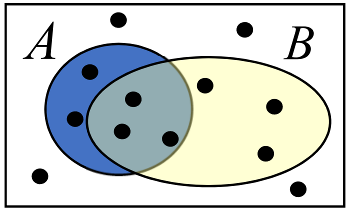
$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$$

- **Normalization**
  - The probability of the entire sample space Ω is equal to 1, that is,

$$\mathbf{P}(\Omega) = 1$$

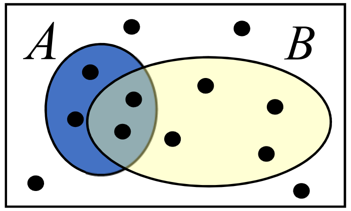
Conditional Probability (1/2)

- Conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information
  - Suppose that the outcome is within some given event B, we wish to quantify the likelihood that the outcome also belongs some other given event A



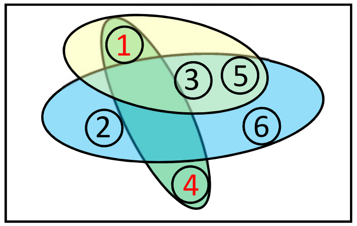
  - Using a new probability law, we have the **conditional probability of** A **given** B, denoted by $\mathbf{P}(A|B)$, which is defined as:

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



    - If $\mathbf{P}(B)$ has zero probability, $\mathbf{P}(A|B)$ is undefined
    - We can think of $\mathbf{P}(A|B)$ as out of the total probability of the elements of B, the fraction that is assigned to possible outcomes that also belong to A

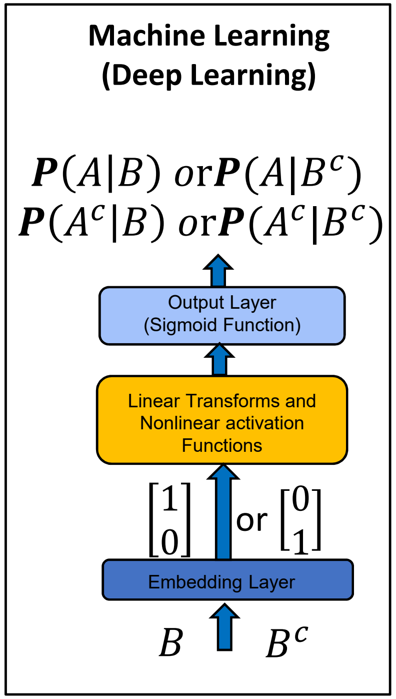
Conditional Probability (2/2)



- When all outcomes of the experiment are equally likely, the conditional probability also can be defined as

$$\mathbf{P}(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}$$

- Some examples having to do with conditional probability
  1. In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?
  2. In a word guessing game, the first letter of the word is a "*t*". What is the likelihood that the second letter is an "*h*"?
  3. How likely is it that a person has a disease given that a medical test was negative?
  4. A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?



Machine Learning (Deep Learning)

$\boldsymbol{P}(A|B)$ or $\boldsymbol{P}(A|B^c)$
$\boldsymbol{P}(A^c|B)$ or $\boldsymbol{P}(A^c|B^c)$

Output Layer (Sigmoid Function)

Linear Transforms and Nonlinear activation Functions

$\begin{bmatrix}1\\0\end{bmatrix}$ or $\begin{bmatrix}0\\1\end{bmatrix}$

Embedding Layer

B B^c

Multiplication (Chain) Rule

- Assuming that all of the conditioning events have positive probability, we have

$$\mathbf{P}(\cap_1^n A_i) = \mathbf{P}(A_1)\mathbf{P}\big(A_2\big|A_1\big)\mathbf{P}\big(A_3\big|A_1 \cap A_2\big)\cdots\mathbf{P}(A_n| \cap_1^{n-1} A_i)$$

  - The above formula can be verified by writing

$$\mathbf{P}(\cap_1^n A_i) = \mathbf{P}(A_1)\frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)}\frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)}\cdots\frac{\mathbf{P}(\bigcap_{i=1}^n A_i)}{\mathbf{P}\big(\bigcap_{i=1}^{n-1} A_i\big)}$$

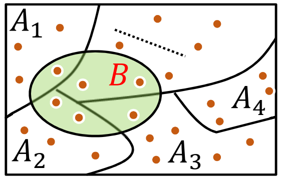
  - For the case of just two events, the multiplication rule is simply the definition of conditional probability

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}\big(A_2\big|A_1\big)$$

Total Probability Theorem (1/2)

- Let $A_1, \cdots, A_n$ be disjoint events that form a partition of the sample space and assume that $\mathbf{P}(A_1) > 0$, for all i . Then, for any event B , we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}\left(B \middle| A_1\right) + \cdots + \mathbf{P}(A_n)\mathbf{P}\left(B \middle| A_n\right)\end{aligned}$$



  - Note that each possible outcome of the experiment (sample space) is included in one and only one of the events $A_1, \cdots, A_n$

Calculate the probability of an event in a divide-and-conquer manner.

Total Probability Theorem (2/2)

Figure 1.13: Visualization and verification of the total probability theorem. The events $A_1, \ldots, A_n$ form a partition of the sample space, so the event B can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets A_i, i.e.,

$$B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B).$$

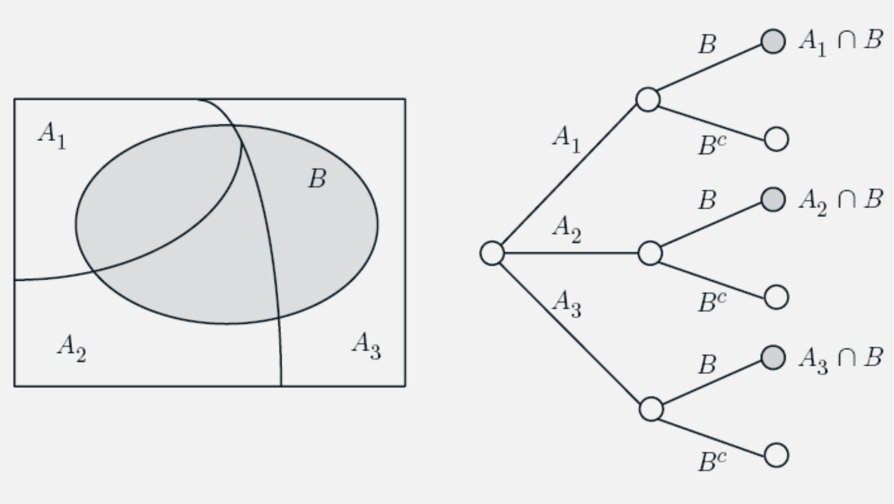
Using the additivity axiom, it follows that

$$\mathbf{P}(B) = \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B).$$

Since, by the definition of conditional probability, we have

$$\mathbf{P}(A_i \cap B) = \mathbf{P}(A_i)\mathbf{P}(B \mid A_i),$$

the preceding equality yields

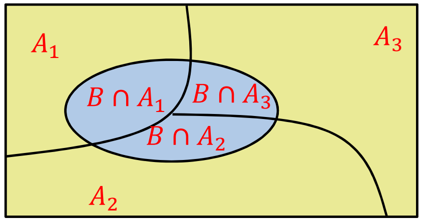


$$\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n).$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability of the leaf $A_i \cap B$ is the product $\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)$ of the probabilities along the path leading to that leaf. The event B consists of the three highlighted leaves and $\mathbf{P}(B)$ is obtained by adding their probabilities.

Bayes' Rule

- Let $A_1, A_2, \dots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$, for all i . Then, for any event B such that $\mathbf{P}(B) > 0$ we have



$\mathbf{P}(A_i)$: an initial belief of a cenario A_i (e.g., a dice's points)

$\mathbf{P}(A_i|B)$: a revised belief of a cenario A_i given that B happens

(e.g., the colors, black and red, on the dice's points)

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)}$$

Multiplication rule

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(B)}$$

Total probability theorem

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\sum_{k=1}^{n} \mathbf{P}(A_k)\mathbf{P}(B|A_k)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\mathbf{P}(A_1)\mathbf{P}(B|A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B|A_n)}$$

Bayes’ Rule and Inference

- Put forwarded by Thomas Bayes (c. 1701-1761), Presbyterian minister
- “Bayes’ theorem,” published posthumously
- A systematic approach for learning from experience and incorporating new evidence

- Bayesian Inference
  - Initial beliefs $\mathbf{P}(A_i)$ on possible causes of an observed event B
  - Establish a model of the world given each A_i: $\mathbf{P}(B|A_i)$

$$A_i \xrightarrow[\mathbf{P}(B|A_i)]{\text{model}} B$$

  - Drawn conclusion about causes

$$B \xrightarrow[\mathbf{P}(A_i|B)]{\text{inference}} A$$

Independence (1/2)

- Recall that conditional probability $\mathbf{P}(A|B)$ captures the partial information that event B provides about event A

- A special case arises when the occurrence of B provides no such information and does not alter the probability that A has occurred

$$\mathbf{P}(A|B) = \mathbf{P}(A)$$

  - A is independent of B (B also is independent of A)

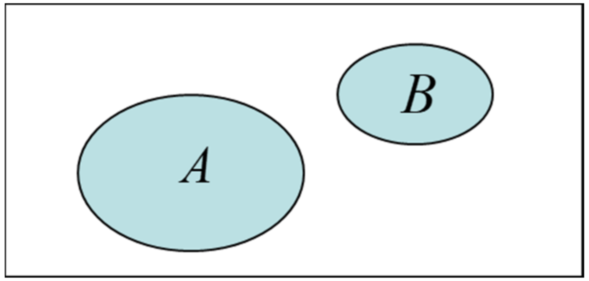
$$\Rightarrow \mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \mathbf{P}(A)$$

$$\Rightarrow \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$$

Independence (2/2)

- A and B are independent => A and B are disjoint (?)
  - No ! Why ?
    - A and B are disjoint then $\mathbf{P}(A \cap B) = 0$
    - However, if $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$

$$\Rightarrow \mathbf{P}(A \cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$$



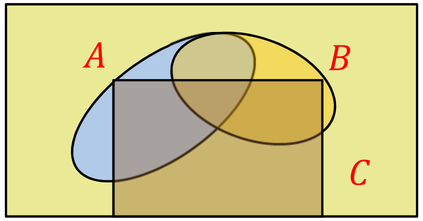
- Two disjoint events A and B with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$ are never independent
- Any event and the event with no outcome (i.e., the empty event) are independent of each other (?)
- Any event and its complement are not independent of each other (?)

Conditional Independence (1/2)

- Given an event C, the events A and B are called conditionally independent if

$$\mathbf{P}(A \cap B|C) = \boxed{\mathbf{P}(A|C)\mathbf{P}(B|C)} \quad 1$$

  - We also know that



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

multiplication rule

$$= = \frac{\cancel{\mathbf{P}(C)}\boxed{\mathbf{P}(B|C)\mathbf{P}(A|B \cap C)}}{\cancel{\mathbf{P}(C)}} \quad 2$$

  - If $\mathbf{P}(B|C) > 0$, we have an alternative way to express conditional independence

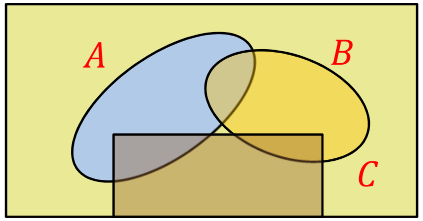
$$\mathbf{P}(A|B \cap C) = \mathbf{P}(A|C) \quad 3$$

Conditional Independence (2/2)

- Notice that independence of two events A and B with respect to the unconditionally probability law does not imply conditional independence, and vice versa

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \quad \xcancel{\Leftrightarrow} \quad \mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$$

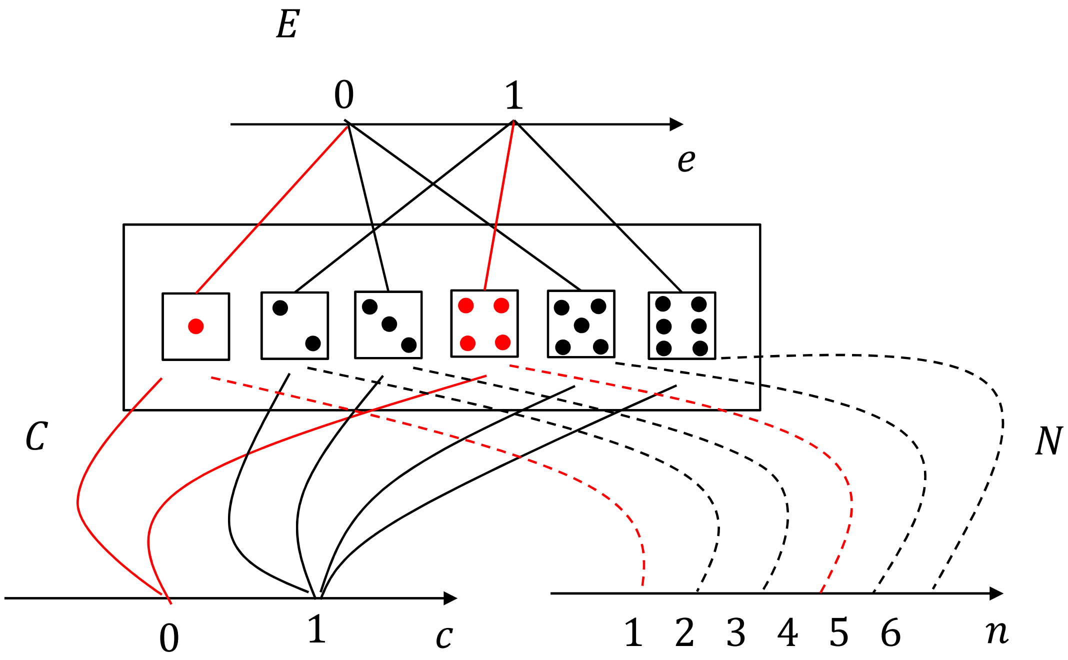
If A and B are independent,



Are A and B independent given that C occured?

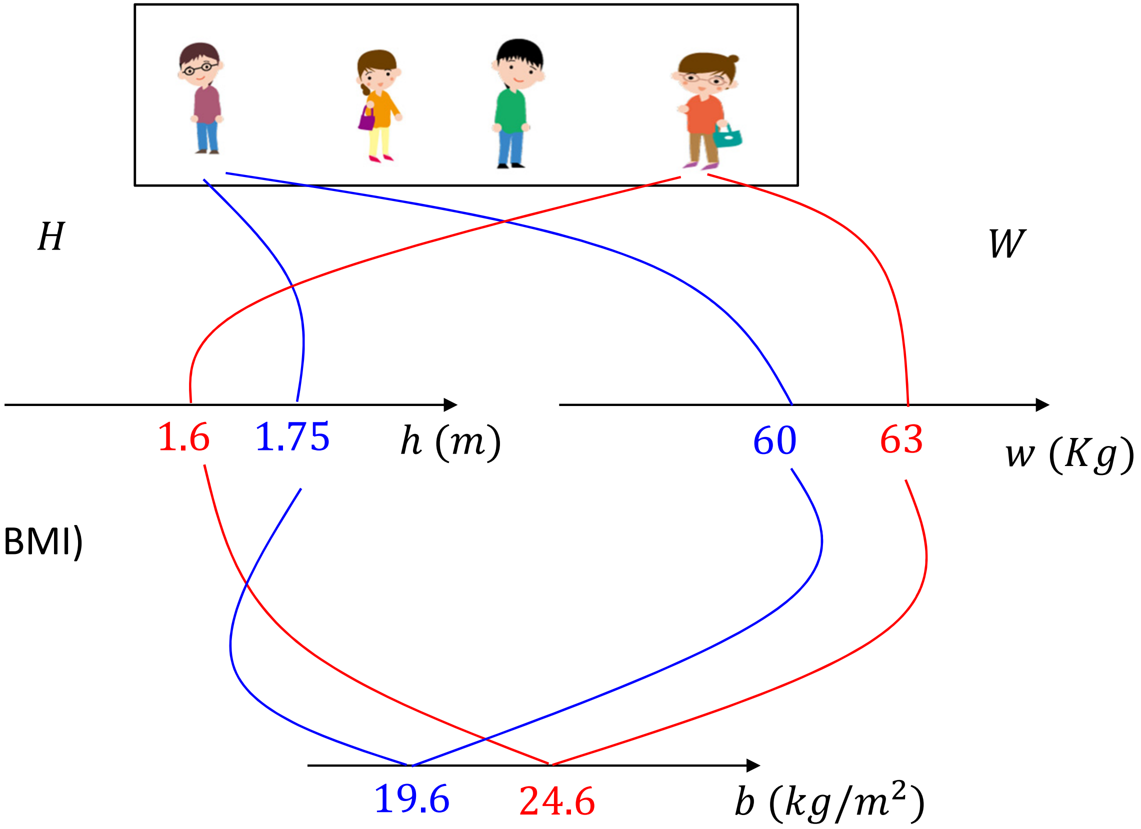
Notion of Random Variables (1/2)

- An experiment consists a roll of a six-sided die



Notion of Random Variables (2/2)

- An experiment consists of a *m*-person population

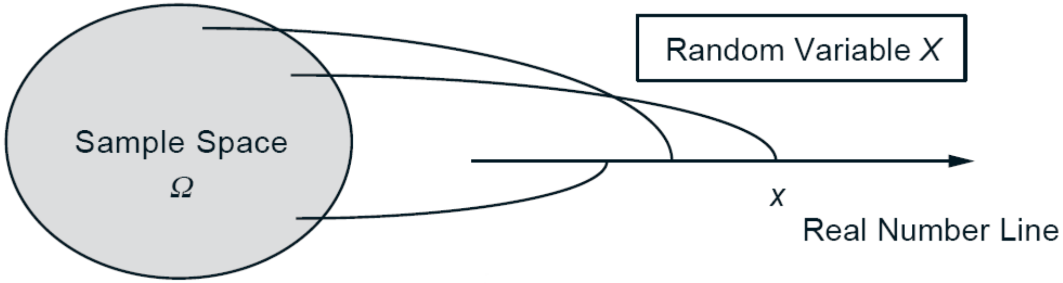


Body Mass Index (BMI)

$$B = \frac{W}{H^2}$$

Random Variables

- Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome
  - This number is referred to as the (numerical) value of the random variable
  - We can say a random variable is a real-valued function of the experimental outcome



Discrete/Continuous Random Variables (1/2)

- A random variable is called **discrete** if its **range** (the set of values that it can take) is finite or at most countably infinite

$$\text{finite}: \{1, 2, 3, 4\}, \text{countably infinite}: \{1, 2, \cdots\}$$

- A random variable is called **continuous (not discrete)** if its **range** (the set of values that it can take) is uncountably infinite
  - E.g., the experiment of choosing a point a from the interval $[-1, 1]$
    - A random variable that associates the numerical value a^2 to the outcome a is not discrete

Discrete/Continuous Random Variables (2/2)

- A discrete random variable X has an associated **probability mass function** (PMF), $p_X(x)$, which gives the probability of each numerical value that the random variable can take

- A continuous random variable X can be described in terms of a **nonnegative** function $f_X(x)$ $(f_X(x) \geq 0)$, called the **probability density function** (PDF) of X , which satisfies

for every subset *B* of the real line $\mathbf{P}(X \in B) = \int_B f_X(x)dx$

Interpretation of PDF

- For an interval $[x, x+\delta]$ with very small length δ, we have

$$P([x, x+\delta]) = \int_x^{x+\delta} f_X(t)dt \approx f_X(x)\cdot\delta$$

  - Therefore, $f_X(x)$ can be viewed as the "probability mass per unit length" near x

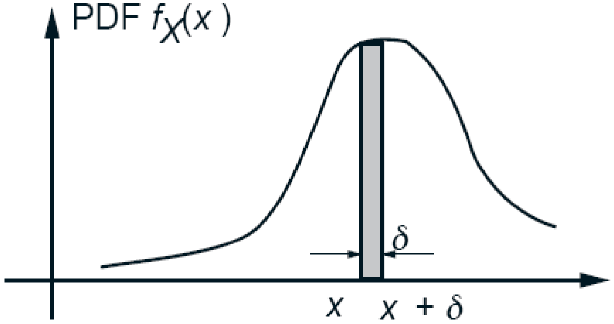


Figure 3.2: Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around x. If δ is very small, the probability that X takes value in the interval $[x, x+\delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x)\cdot\delta$.

- $f_X(x)$ is not the probability of any particular event, it is also not restricted to be less than or equal to one

Cumulative Distribution Functions (1/4)

- The cumulative distribution function (CDF) of a random variable X is denoted by $F_X(x)$ and provides the probability $\mathbf{P}(X \leq x)$

$$F_X(x) = \mathbf{P}(X \leq x) = \begin{cases} \displaystyle\sum_{k \leq x} p_X(k)\,, & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{x} f_X(t)dt\,, & \text{if } X \text{ is continuous} \end{cases}$$

  - The CDF $F_X(x)$ accumulates probability up to x
  - The CDF $F_X(x)$ provides a unified way to describe all kinds of random variables mathematically

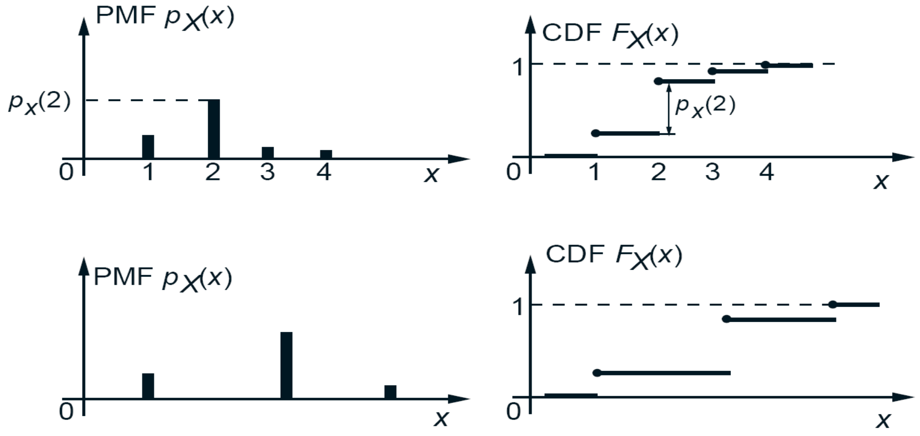
Cumulative Distribution Functions (2/4)

- The CDF $F_X(x)$ is monotonically non-decreasing

$$\text{if } x_i \leq x_j, \text{ then } F_X(x_i) \leq F_X(x_j)$$

- The CDF $F_X(x)$ tends to 0 as $x \to -\infty$, and to 1 as $x \to \infty$

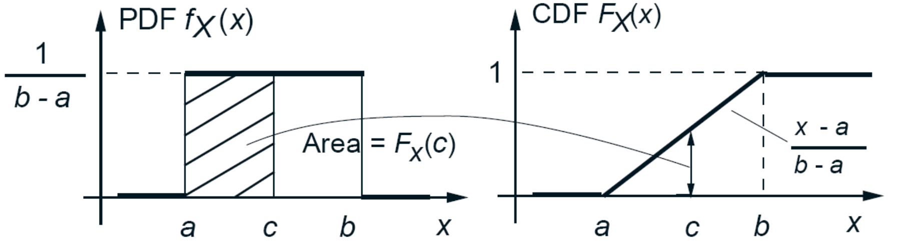
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x



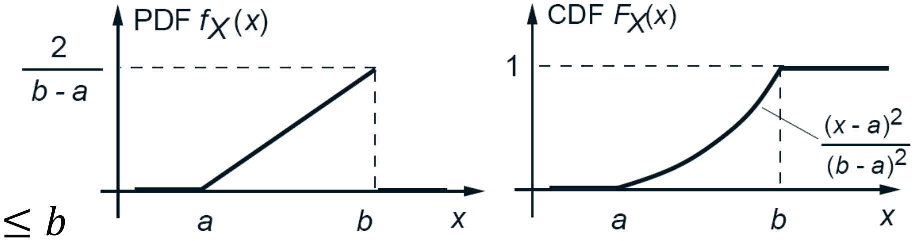
Cumulative Distribution Functions (3/4)

- If X is continuous, then $F_X(x)$ is a continuous function of x

$$f_X(x) = \frac{1}{b-a},\ for\ a \le x \le b$$



$$F_x(X \le x) = \int_a^x f_X(t)\,dt$$
$$= \int_a^x \frac{1}{b-a}\,dt$$
$$= \frac{x-a}{b-a}$$



$$f_X(x) = c(x-a), \qquad for\ a \le x \le b$$
$$\Rightarrow \int_a^b c(x-a)dx = \frac{c}{2}(x-a)^2 \Big|_a^b = 1$$
$$\Rightarrow c = \frac{2}{(b-a)^2}$$
$$\Rightarrow f_X(b) = \frac{2(b-a)}{(b-a)^2} = \frac{2}{b-a}$$

$$F_x(X \le x) = \int_a^x f_X(t)\,dt$$
$$= \int_a^x \frac{2(t-a)}{(b-a)^2}dt = \frac{(x-a)^2}{(b-a)^2}$$

Cumulative Distribution Functions (4/4)

- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing

$$F_X(k) = \mathbf{P}(X \le k) = \sum_{i=-\infty}^{k} p_X(i)\,,$$
$$p_X(k) = \mathbf{P}(X \le k) - \mathbf{P}(X \le k-1) = F_X(k) - F_X(k-1)$$

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation

$$F_X(x) = \mathbf{P}(X \le x) = \int_{-\infty}^{x} f_X(t)dt\,,$$
$$f_X(x) = \frac{dF_X(x)}{dx}$$

  - The second equality is valid for those x for which the CDF has a derivative (e.g., the piecewise constant random variable)

Conditioning

- Let X and Y be two random variables associated with the same experiment
  - If X and Y are discrete, the conditional PMF of X is defined as (where $p_Y(y)$)

$$p_{X|Y}(x|y) = \mathbf{P}(X = x|Y = y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

  - If X and Y are continuous, the conditional PDF of X is defined as (where $f_Y(y) > 0$)

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Independence

- Two **random variables** X and Y are **independent** if

$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x, y$ (If X and Y are discrete)

$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for all } x, y$ (If X and Y are continuous)

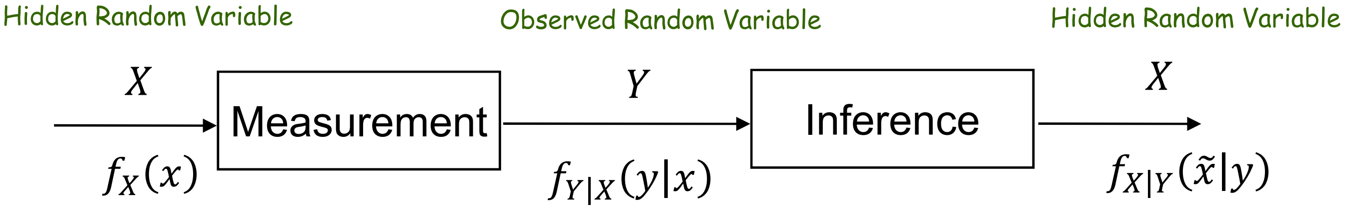
- If two **random variables** X and Y are **independent**

$p_{X|Y}(x|y) = p_X(x), \quad \text{for all } x, y$ (If X and Y are discrete)

$f_{X|Y}(x|y) = f_X(x), \quad \text{for all } x, y$ (If X and Y are continuous)

Inference and the Continuous Bayes' Rule (1/2)

- As we have a model of an underlying but unobserved phenomenon, represented by a random variable X with PDF f_X, and we make a noisy measurement Y, which is modeled in terms of a conditional PDF $f_{Y|X}$. Once the experimental value of Y is measured, what information does this provide on the unknown value of X? (the so-called noisy channel model)



$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t) f_{Y|X}(y|t) dt}$$

Note that we have

$$f_X f_{Y|X} = f_{X,Y} = f_Y f_{X|Y}$$

Inference and the Continuous Bayes' Rule (2/2)

Inference about a Discrete Random Variable

- If the unobserved phenomenon is inherently discrete
  - Let N is a discrete random variable of the form $\{N = n\}$ that represents the different discrete probabilities for the unobserved phenomenon of interest, and p_N be the PMF of N

$$\mathbf{P}(N = n|Y = y) \approx \mathbf{P}(N = n|y \le Y \le y + \delta)$$

$$= \frac{\mathbf{P}(N = n)\mathbf{P}(y \le Y \le y + \delta|N = n)}{\mathbf{P}(y \le Y \le y + \delta)}$$

$$\approx \frac{p_N(n)f_{Y|N}(y|n)\cancel{\delta}}{f_Y(y)\cancel{\delta}}$$

$$\approx \frac{p_N(n)f_{Y|N}(y|n)}{f_Y(y)}$$

$$= \frac{p_N(n)f_{Y|N}(y|n)}{\sum_i p_N(i)f_{Y|N}(y|i)}$$

Total probability theorem

33

Inference Based on a Discrete Random Variable

- The earlier formula expressing $\mathbf{P}(A|Y=y)$ in terms of $f_{Y|A}(y)$, which can be turned around to yield

(See pp. 33 of this handout)

$$f_{Y|A}(y) \overset{?}{\simeq} \frac{f_Y(y)\mathbf{P}(A|Y=y)}{\mathbf{P}(A)}$$

?

$$= \frac{f_Y(y)\mathbf{P}(A|Y=y)}{\int_{-\infty}^{\infty} f_Y(t)\mathbf{P}(A|Y=t)dt}$$

$$\mathbf{P}(A)f_{Y|A}(y) = f_Y(y)\mathbf{P}(A|Y=y)$$

$$\Rightarrow \int_{-\infty}^{\infty} \mathbf{P}(A)f_{Y|A}(y)dy = \int_{-\infty}^{\infty} f_Y(y)\mathbf{P}(A|Y=y)dy$$

$$\Rightarrow \mathbf{P}(A) = \int_{-\infty}^{\infty} f_Y(y)\mathbf{P}(A|Y=y)dy \ \ (\because \text{normalization property: } \int_{-\infty}^{\infty} f_{Y|A}(y)dy = 1)$$

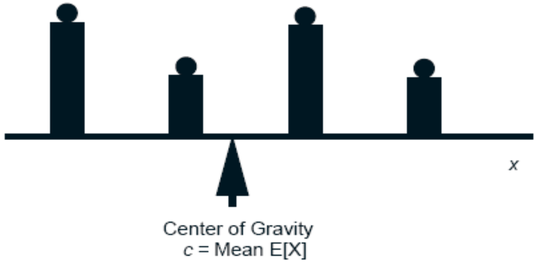
Discrete Random Variables: Expectation

- The **expected value** (also called the **expectation** or the **mean**) of a discrete random variable X, with PMF p_X, is defined by

$$\mathbf{E}[X] = \sum_x x p_X(x)$$

  - Can be interpreted as the **center of gravity** of the PMF (Or a weighted average, in proportion to probabilities, of the possible values of X)

- The expectation is well-defined if



$$\sum_x (x-c)p_X(x) = 0$$

$$\Rightarrow c = \sum_x x \cdot p_X(x)$$

$$\sum_x |x| p_X(x) < \infty$$

  - That is, $\sum_x x p_X(x)$ converges to a finite value

Discrete Random Variables: Moments

- The ***n*-th moment** of a discrete random variable X is the expected value of a random variable X^n (or the random variable Y , $Y = g(X) = X^n$)

$$\mathbf{E}[X^n] \overset{?}{=} \sum_x x^n p_X(x)$$

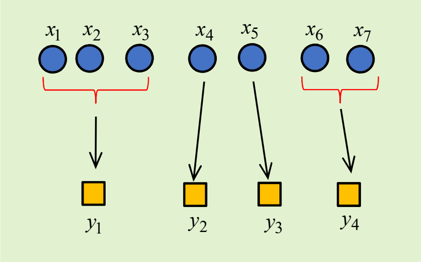
  - The 1st moment of a random variable X is just its mean (or expectation)

X^n is termed as *X* raised to the power of *n* (or the *n*th power), or the *n*th power of *X*.

Expectations for Functions of Discrete Random Variables

- Let X be a random variable with PMF p_X, and let $g(X)$ be a function of X. Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x)$$



- To verify the above rule
  - Let $Y = g(X)$, and therefore $p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x)$

$$\mathbf{E}[g(X)] = \mathbf{E}[Y] = \sum_y y\, p_Y(y)$$

$$= \sum_y y \sum_{\{x|g(x)=y\}} p_X(x) = \boxed{\sum_y \sum_{\{x|g(x)=y\}}} g(x) p_X(x)$$

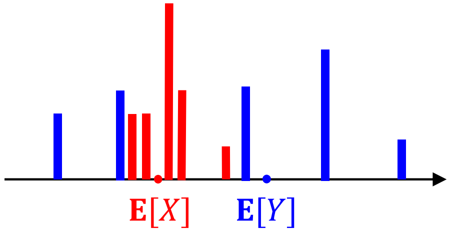
$$= \boxed{\sum_x} g(x) p_X(x)$$

?

Discrete Random Variables: Variance

- The **variance** of a random variable X is the expected value of a random variable $\left(X - \mathbf{E}(X)\right)^2$

$$\mathrm{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$$
$$\overset{?}{=} \sum_x (x - \mathbf{E}[X])^2 p_X(x)$$



  - The variance is always nonnegative (why?)
  - The variance provides a measure of dispersion of X around its mean
  - The standard derivation is another measure of dispersion, which is defined as (a square root of variance)

$$\sigma_X = \sqrt{\mathrm{var}(X)}$$

    - Easier to interpret, because it has the same units as X and capture the width of X's distribution

Variance in Terms of Moments Expression

- We can also express variance of a random variable X as

$$\mathrm{var}(X) = \underbrace{\mathbf{E}[X^2]}_{\text{Second Moment}} - \underbrace{(\mathbf{E}[X])^2}_{\text{Square of First Moment}}$$

$$\begin{aligned}
\mathrm{var}(X) &= \sum_x (x - \mathbf{E}[X])^2 \, p_X(x) \\
&= \sum_x (x^2 - 2x\mathbf{E}[X] + (\mathbf{E}[X])^2) \, p_X(x) \\
&= \left[\sum_x x^2 p_X(x)\right] + 2\mathbf{E}[X]\left[\sum_x x p_X(x)\right] + (\mathbf{E}[X])^2\left[\sum_x p_X(x)\right] \\
&= \mathbf{E}[X^2] - 2(\mathbf{E}[X])^2 + (\mathbf{E}[X])^2 \\
&= \mathbf{E}[X^2] - (\mathbf{E}[X])^2
\end{aligned}$$

More on Expectation and Moments (1/2)

- The **expectation** of a random variable X is defined by

$$\mathbf{E}[X] = \sum_x x p_X(x)$$ (If X is discrete)

or

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$ (If X is continuous)

- The ***n*-th moment** of a random variable X is the expected value of a random variable X^n (or the random variable

$$\mathbf{E}[X^n] = \sum_x x^n p_X(x)$$ (If X is discrete)

or

$$\mathbf{E}[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$ (If X is continuous)

  - The 1st moment of a random variable is just its mean

More on Expectation and Moments (2/2)

- Let X be a random variable and let $Y = aX + b$

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b$$

$$\mathrm{var}(Y) = a^2\,\mathrm{var}(X)$$

- If X and Y are independent random variables

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$$

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

g and h are functions of X and Y, respectively

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\mathbf{E}[h(Y)]$$

Variance of the Sum of Two “Dependent” Random Variables

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y)$$

$$\begin{aligned}
&\mathrm{var}(X+Y) = \mathbf{E}[((X+Y) - \mathbf{E}[X+Y])^2] \\
&= \mathbf{E}[(X+Y)^2] - 2\mathbf{E}[X+Y]\,\mathbf{E}[X+Y] + (\mathbf{E}[X+Y]))^2 \\
&= \mathbf{E}[X^2] + \mathbf{E}[Y^2] + 2\mathbf{E}[XY] - \mathbf{E}[X+Y]\,\mathbf{E}[X+Y] \\
&= \mathbf{E}[X^2] + \mathbf{E}[Y^2] + 2\mathbf{E}[XY] - (\mathbf{E}[X] + \mathbf{E}[Y])^2 \\
&= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 + \mathbf{E}[Y^2] - (\mathbf{E}[Y])^2 + 2(\mathbf{E}[XY] - \mathbf{E}[X]E[Y]) \\
&= \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y)
\end{aligned}$$

Note that $\mathrm{cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$

$$= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[X]$$

Two Useful Probability Laws

- **Law of Iterated Expectations**

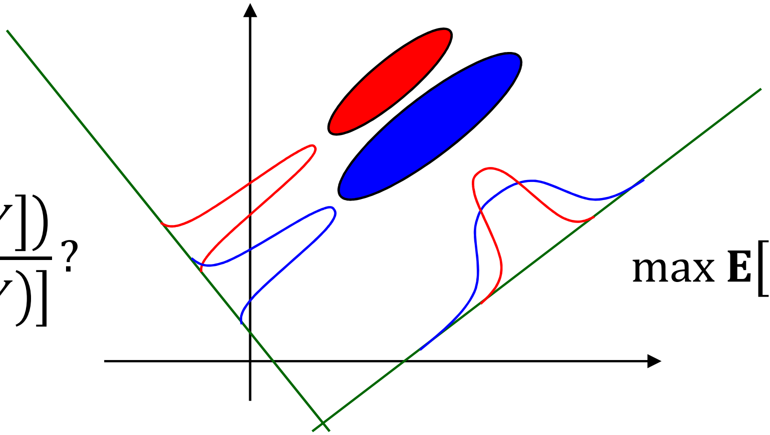
$$\mathbf{E}\big[\mathbf{E}[X|Y]\big] = \mathbf{E}[X]$$

- **Law of Total Variance**

$$\operatorname{var}(X) = \mathbf{E}[\operatorname{var}(X|Y)] + \operatorname{var}(\mathbf{E}[X|Y])$$

LDA: Linear Discriminant Analysis

$$\max \frac{\operatorname{var}\big(\mathbf{E}[\vec{V}^T\vec{X}|Y]\big)}{\mathbf{E}\big[\operatorname{var}(\vec{V}^T\vec{X}|Y)\big]}?$$



PCA: Principal Component Analysis

$$\max \mathbf{E}\big[\operatorname{var}(\vec{V}^T\vec{X}|Y)\big] + \operatorname{var}\big(\mathbf{E}[\vec{V}^T\vec{X}|Y]\big)?$$

More on Conditional Expectation

$$\mathbf{E}[X] = \mathbf{E}\big[\mathbf{E}[X|Y]\big] = \begin{cases} \sum_y \mathbf{E}[X|Y=y]p_Y(y) & \text{(if } Y \text{ is discrete)} \\ \int_{-\infty}^{\infty} \mathbf{E}[X|Y=y]f_Y(y)dy & \text{(if } Y \text{ is continuous)} \end{cases}$$

$$\mathbf{E}[X|Y=y] = \begin{cases} \sum_x x p_{X|Y}(x|y) & \text{(if } X \text{ is discrete)} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx & \text{(if } X \text{ is continuous)} \end{cases}$$

Bayesian Statistics (1/6)

- Frequentist Statistics vs. Bayesian Statistics
- Bayesian updating
  - A coin is tossed 10 times and gets 8 heads
    - This coin comes down heads 8 times out of 10 (from a frequentist point of view)
    - This is the Maximum Likelihood Estimate (MLE)

$$P(\text{head}) = 0.8$$

$$P(\text{tail}) = 0.2$$

  - Bayesian statistics: measure degree of belief, and are calculated by starting with prior belief and updating them in the face of evidence, by use of Bayes' theorem

Bayesian Statistics (2/6)

- Let θ_m be the model that asserts $P(\text{head}) = m$, ***s*** be a sequence of observations, i heads and j tails
  - For any m, $0 \leq m \leq 1$:

$$P(\mathbf{s}|\theta_m) = m^i(1-m)^j$$

- From a frequentist point of view, we wish to find the MLE

$$\underset{m}{\arg\max}\, P(\mathbf{s}|\theta_m) \overset{?}{=} \underset{m}{\arg\max} \log P(\mathbf{s}|\theta_m)$$

  - logarithmic functions are monotone increasing functions
  - We can differentiate the above polynomial then the answer is $\frac{i}{i+j}$, or 0.8 for the case of 8 heads and 2 tails

Bayesian Statistics (3/6)

- Bayesian Updating:
  - Let us instead assume that one's prior belief is modeled by the distribution

$$f(\theta_m) = 6m(1-m)$$

    - This polynomial was chosen because its distribution is centered on 1/2, and, conveniently, the area under the curve between 0 and 1 is 1

  - When one sees an observation sequence s one wants to know one's new belief in the fairness of the coin. By Bayes' theorem

$$f(\theta_m|\mathbf{s}) = \frac{P(\mathbf{s}|\theta_m)f(\theta_m)}{P(\mathbf{s})} = \frac{m^i(1-m)^j \times 6m(1-m)}{P(\mathbf{s})} = \frac{6m^{i+1}(1-m)^{j+1}}{P(\mathbf{s})}$$

Bayesian Statistics (4/6)

- $P(\mathbf{s})$ is the prior probability of $\mathbf{s},$ and we can ignore it while finding the m that maximizes the above equation ($P(\mathbf{s})$ is a normalization factor)

- If we then differentiate the numerator so as find its maximum, we can determine that for the case of 8 heads and 2 tails:

$$\arg\max_{m} f(\theta_m|\mathbf{s}) = 3/4$$

the maximum of the a posteriori distribution (MAP)

  - Because our prior was weak (the polynomial is a quite flat curve centered over $1/2$), we have moved a long way in the direction of believing that the coin is biased, but the important point is that we haven't moved all the way to 0.8
  - If we had assumed a stronger prior, we would have moved a smaller distance from $1/2$

Bayesian Statistics (5/6)

$$\underset{m}{\arg\max}\, f(\theta_m|\mathbf{s}) = \underset{m}{\arg\max}\, \frac{6m^{i+1}(1-m)^{j+1}}{P(\mathbf{s})}$$

$$\propto \underset{m}{\arg\max}\, 6m^{i+1}(1-m)^{j+1}$$

$$\overset{?}{=} \underset{m}{\arg\max}\, \log 6m^{i+1}(1-m)^{j+1}$$

$$\frac{\partial \log 6m^{i+1}(1-m)^{j+1}}{\partial m} = \frac{m}{i+1} - \frac{1-m}{j+1}$$

$$\frac{\partial \log 6m^{i+1}(1-m)^{j+1}}{\partial m} = 0 \Rightarrow m = \frac{i+1}{i+j+2}$$

$$\therefore \underset{m}{\arg\max}\, P(\theta_m|\mathbf{s}) = \frac{3}{4}$$

Bayesian Statistics (6/6)

- More on $P(\mathbf{s})$
  - This marginal probability which can be obtained by taking integral of all the $P(\mathbf{s}|\theta_m)$ weighted by the probability of $f(\theta_m)$

$$P(\mathbf{s}) = \int_0^1 P(\mathbf{s}|\theta_m)f(\theta_m)dm$$
$$= \int_0^1 6m^{i+1}(1-m)^{j+1}dm$$

  - This just happens to be an instance of the Beta integral, another continuous distribution well-studied by statisticians. we can look up to find out that

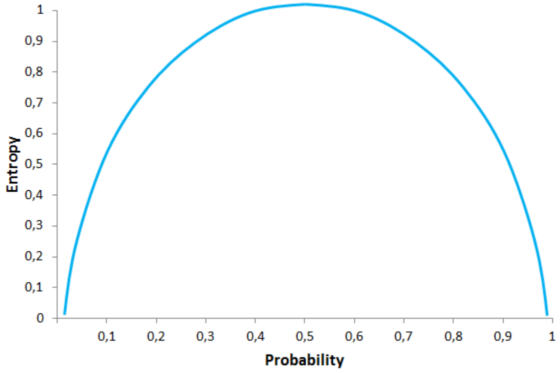
$$P(\mathbf{s}) = \frac{6(i+1)!(j+1)!}{(i+J+3)!}$$

Entropy (1/3)

- Entropy measures the amount of information in a random variable

$$H(X) = -\sum_{x \in X} p(x) \log_2 p\,(x) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} = \mathbf{E}[\log_2 \frac{1}{p(X)}]$$

- We define $0 \log_2 0 = 0$



The entropy of a weighted coin. The horizontal axis shows the probability of a weighted coin to come up heads. The vertical axis shows the entropy of tossing the corresponding coin once.

- Entropy can be regarded as
  - The average uncertainty of a single random variable
  - The average length of the message needed to transmit an outcome of that variable
  - We can think of entropy as a matter of how surprised we will be
  - We hope the entropy is lower in the system (?)

Entropy (2/3)

- Example: Suppose you are reporting the result of rolling an fair 8-sided die
  - Then the entropy is

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{8} p(i)\log_2 p(i) \\
&= -\textstyle\sum_{i=1}^{8} \frac{1}{8}\log_2\frac{1}{8} \\
&= -\log_2\frac{1}{8} \\
&= \log_2 8 \\
&= 3
\end{aligned}
$$

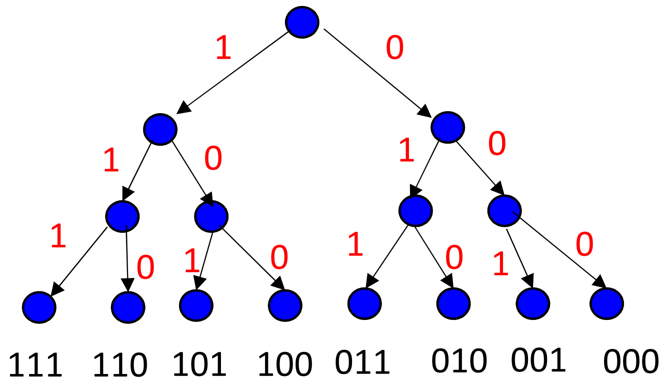
  - The most efficient way is to simply encode the result as a 3 digit binary message

Result:	1	2	3	4	5	6	7	8
Binary Encoding:	001	010	011	100	101	110	111	000

Entropy (3/3)

- Entropy can be interpreted as a measure of the size of the “search space” consisting of the possible values of a random variable and its associated probabilities (?)

- Note that:
  - $H(X) \geq 0$
  - $H(X) = 0$ only when the value of X is determinate (providing no new information)
  - Entropy increases with the message length



- Another example: simplified Polynesian language with six letters

Letter:	p	t	k	a	i	u
Probability:	1/8	1/4	1/8	1/4	1/8	1/8
Binary Encoding:	100	00	101	01	110	111

$$H(X) = 2\frac{1}{2} \text{ (bits)}$$

Joint Entropy and Conditional Entropy (1/2)

- Joint Entropy
  - The amount of information needed on average to specify both their values

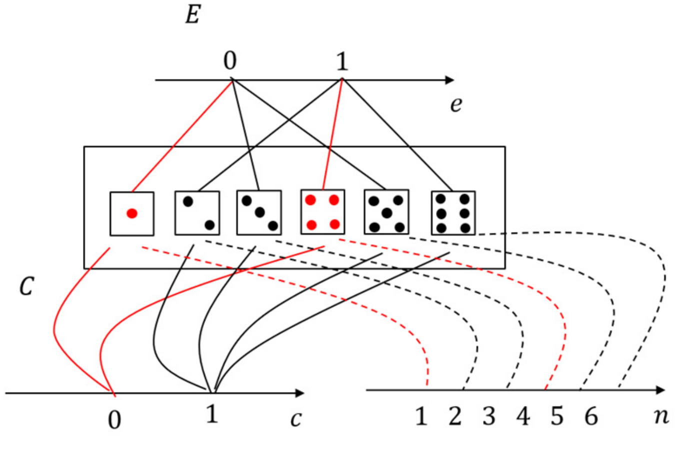
$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p\,(x,y) = \mathrm{E}[\log\frac{1}{p(X,Y)}]$$

- Conditional Entropy:
  - How much extra information you still need to supply on average to communicate Y given that the other party knows X

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

$$= \sum_{x\in X} p(x)\left[-\sum_{y\in Y} p(y|x)\log p\,(y|x)\right]$$

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



Joint Entropy and Conditional Entropy (2/2)

- Chain Rule for Entropy

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

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_2|X_1, \dots, X_{n-1})$$

- Proof:

$$\boxed{\because p(x,y) = p(y|x)p(x)}$$

$$\begin{aligned}
H(X,Y) &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p\,(x,y) \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log[p(y|x)p(x)] \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x,y)[\log p\,(y|x) + \log p\,(x)] \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p\,(y|x) - \sum_{x \in X} \underbrace{\boxed{\sum_{y \in Y} p(x,y)}}_{p(x)} \log p\,(x) \\
&= H(Y|X) + H(X)
\end{aligned}$$

Simplified Polynesian Language Revisited

$C_1V_1\ C_2V_2\ C_3V_3 \ldots.\ C_T,V_T$

- Simplified Polynesian has syllable structure, viz. all words consist of sequences of CV (consonant-vowel) syllables (totally, 7 syllables)
- This suggests a better model in terms of two random variables C for the consonant of a syllable, and V for the vowel

?

Letter:	p	t	k	a	i	u
Probability:	1/16	3/8	1/16	1/4	1/8	1/8

Consonants

Vowels

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

The letters have a different probability distribution than the previous example.

$$H(C) = -\sum_{c=p,t,k} p(C=c)\log_2 p(C=c) = 1.061\ (\text{bits})$$

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

$$H(V|C) = -\sum_{c=p,t,k} p(C=c)H(V|C=c)$$

$$= \frac{1}{8}H\left(V_{(\frac{1}{2},\frac{1}{2},0)}|C=p\right) + \frac{3}{4}H\left(V_{(\frac{1}{2},\frac{1}{4},\frac{1}{4})}|C=t\right) + \frac{1}{8}H\left(V_{(\frac{1}{2},0,\frac{1}{2})}|C=k\right)$$

?

$$=1.375\ (\text{bits})$$

$$H(C,V) = 1.061 + 1.375 = 2.436\ (\text{bits})$$

The entropy for whole syllables

Entropy Rate

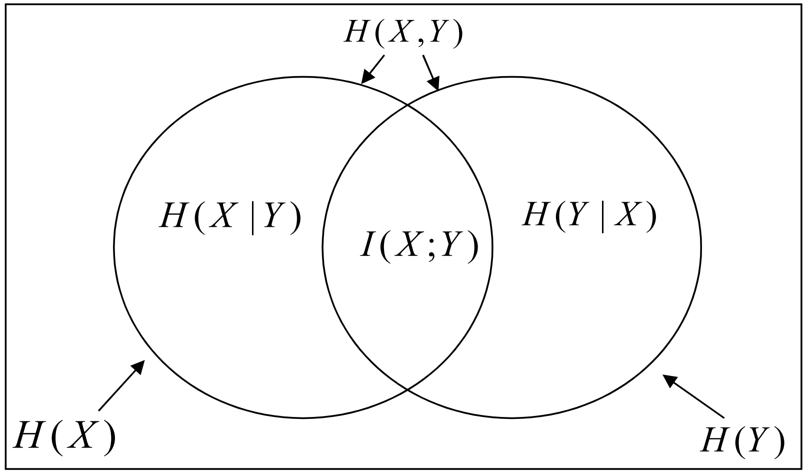
- Because the amount of information contained in a message depends on the length of the message, we normally want to talk in terms of the per-letter or per-word entropy
- For a message of length n, the per-letter or per-word entropy, also known as the entropy rate, is

$$H_{\mathrm{rate}} = \frac{1}{n} H(X_1, X_2, \dots, X_n) = -\frac{1}{n} \sum_{x_{1,} x_2, \dots, x_n} p(x_{1,} x_2, \dots, x_n) \log p\,(x_{1,} x_2, \dots, x_n)$$

Mutual Information (1/5)

- By the chain rule for entropy

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



- Therefore, we have

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

  - This difference is called the mutual information (MI) $I(X;Y)$ between X and Y
  - It is the information reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another

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

Mutual Information (2/5)

- Mutual information is a symmetric, non-negative (?) measure of the common information in the two variables
- It is a measure of independence
  - It is 0 only when two variables are independent
  - For two dependent variables, mutual information grows not only with the degree of dependence, but also according to the entropy of the variables

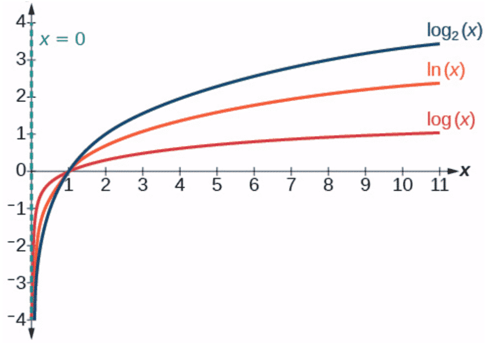
$$H(X,Y) = H(Y) + H(X|Y)$$

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) = \underline{H(X) + H(Y) - H(X,Y)} \\
&= \sum_{x} p(x)\log\frac{1}{p(x)} + \sum_{y} p(y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p\,(x,y) \\
&= \sum_{x,y} p(x,y)\log\frac{1}{p(x)} + \sum_{x,y} p(x,y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p\,(x,y) \\
&= \sum_{x,y} p(x,y)\left[\log\frac{1}{p(x)} + \log\frac{1}{p(y)} + \log p\,(x,y)\right] \\
&= \boxed{\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}}
\end{aligned}$$

$$\therefore I(X;Y) = \mathbf{E}[\log\frac{p(x,y)}{p(x)p(y)}]$$

Mutual Information (3/5)

- Non-negativity Property of Mutual Information

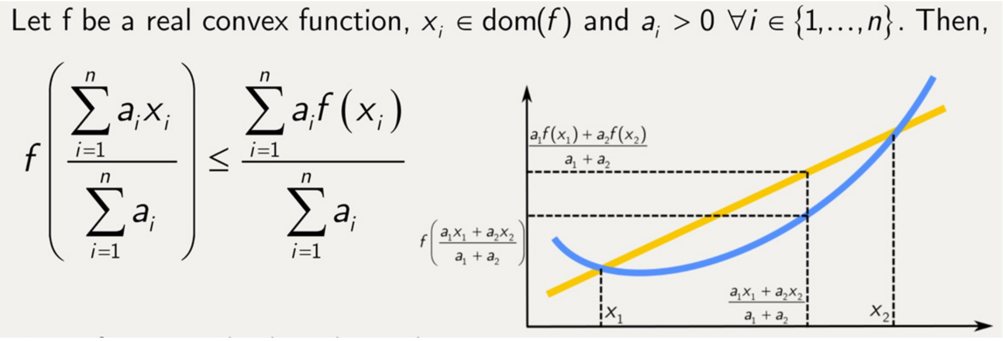


$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$\boxed{\begin{aligned} &= -\textstyle\sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\ &\ge \log \left(\textstyle\sum_{x,y} p(x,y) \frac{p(x)p(y)}{p(x,y)} \right) \end{aligned}}$$

Jensen Inequality for convex functions (negative logarithm is convex)

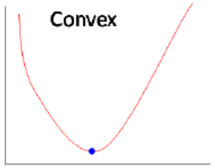
$$\begin{aligned} &= \log\left(\textstyle\sum_{x,y} p(x)p(y) \right) \\ &= \log 1 \\ &= 0 \end{aligned}$$

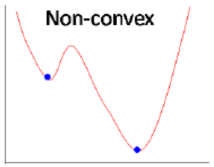


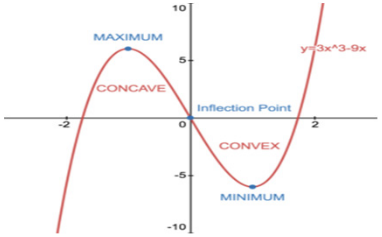
Let f be a real convex function, $x_i \in \mathrm{dom}(f)$ and $a_i > 0\ \forall i \in \{1,\ldots,n\}$. Then,

$$f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} \right) \le \frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i}$$

$\frac{a_1 f(x_1) + a_2 f(x_2)}{a_1 + a_2}$
$f\left(\frac{a_1 x_1 + a_2 x_2}{a_1 + a_2}\right)$
$\frac{a_1 x_1 + a_2 x_2}{a_1 + a_2}$
x_1
x_2







https://twitter.com/mathtype/status/1397191184158691338?lang=zh-Hant

Mutual Information (4/5)

- Properties of Mutual Information
  - $I(X,X) = H(X)$ $\quad (\because H(X)=H(X) - H(\mathrm{X}|\mathrm{X}) = I(X,X),\underset{?}{\underline{\text{where } H(\mathrm{X}|\mathrm{X}) = 0}})$
  - Conditional Mutual Information

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

  - Chain Rule for Mutual Information

    $$I(X_1, X_1, \dots, X_n\,; Y) = I(X_1\,; Y) + \cdots + I(X_n\,; Y|X_1, \dots, X_{n-1})$$

Mutual Information (5/5)

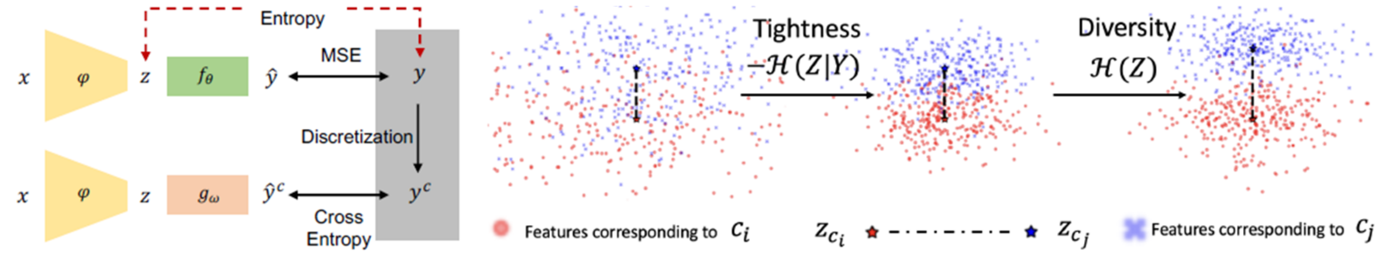
- the pointwise mutual information (PMI) is defined between two particular points

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

  - PMI has sometimes been used as a measure of association between elements, but there are problems with using this measure
  - PMI has been used many times in Statistical NLP, such as for clustering words. It also turns up in word sense disambiguation

More on Mutual Information

- Recall: The mutual information between two random variables can be defined as $I(Z;Y) = H(Z) - H(Z|Y)$ (Z : embeddings, Y : output)
  - An ordinal entropy regularizer is employed to learn highentropy feature representations that preserve ordinality



(a) Illustration of frameworks

(b) Tightness and diversity of ordinal entropy

Figure 2: Illustration of (a) regression and classification for continuous targets, and the use of our ordinal entropy for regression, (b) the pull and push objective of tightness and diversity on the feature space. The tightness part encourages features to be close to their feature centers while the diversity part encourages feature centers to be far away from each other.

S. Zhang et al., “Improving deep regression with ordinal entropy,” ICLR 2023

Kullback-Leibler Divergence (1/3)

- Kullback-Leibler (KL) divergence is also known as Relative Entropy
- For two probability mass functions $p(x)$ and $q(x)$, their relative entropy is given by

$$KL(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = \mathbf{E}_p[\log \frac{p(X)}{q(X)}] \quad (\ 0 \log \frac{0}{q} = 0 \quad \text{and} \quad p \log \frac{p}{0} = \infty \ )$$

- KL divergence a measure of how different two probability distributions (over the same event space) are
  - This quantity is always non-negative (?), also dubbed KL distance, and

$$KL(p||q) = 0 \ \text{ iff } \ p(x) = q(x) \ for \ all \ x$$

  - KL divergence is not symmetric in p and q, $KL(p||q) \neq KL(q||p)$, and it does not satisfy the triangle inequality

$$KL(p||q) \xcancel{\leq} KL(p||h) + KL(h||q)$$

Kullback-Leibler Divergence (2/3)

$$
\begin{aligned}
KL(p||q) &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\
&= -\textstyle\sum_{x \in X} p(x) \log \frac{q(x)}{p(x)} \\
&\geq -\textstyle\sum_{x \in X} p(x) (1 - \frac{q(x)}{p(x)}) \\
&= -(\textstyle\sum_{x \in X} p(x) - \sum_{x \in X} q(x)) \\
&= -(1 - 1) \\
&= 0
\end{aligned}
$$

$$
\boxed{\begin{array}{c} \log a \leq a - 1 \\ \therefore -\log a \geq 1 - a \end{array}}
$$

Kullback-Leibler Divergence (3/3)

- Recall: Mutual information (MI) is actually just a measure of how far a joint distribution is from independence

$$I(X\,;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = KL(p(x,y)||p(x)p(y))$$

- We can also derive conditional relative entropy and a chain rule for relative entropy

$$KL(p(y|x)||q(y|x)) = \sum_{x} p(x)\sum_{y} p(y|x)\log\frac{p(y|x)}{q(y|x)}$$

$$KL(p(x,y)||q(x,y)) \overset{?}{=} KL(p(x)||q(x)) + KL(p(y|x)||q(y|x))$$

Cross Entropy

- The cross entropy (CE) between a random variable X with true probability distribution $p(x)$ and another PMF $q(x)$ (normally a model of p) is given by

$$CE(p,q) = -\sum_{x \in X} p(x) \log q\,(x) = \sum_{x \in X} p(x) \log \frac{1}{q(x)} = \mathbf{E}_p[\log \frac{1}{q(X)}]$$

- Relationship between cross entropy, entropy and KL divergence

different probability models

$$p(x) = P(X = x|M_p)$$

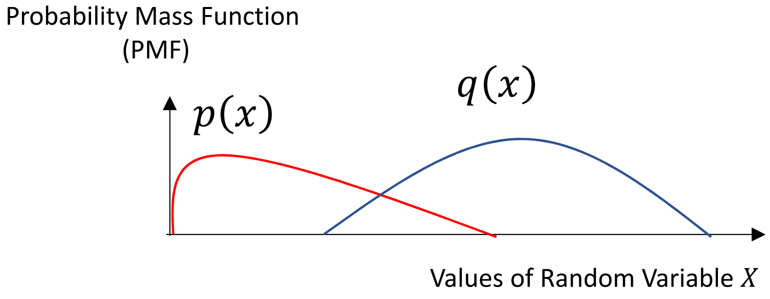
$$q(x) = P(X = x|M_q)$$

$$KL(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_{x \in X} p(x) \log \frac{1}{q(x)} - \sum_{x \in X} p(x) \log \frac{1}{p(x)}$$

$$= CE(p,q) - H^p(X) \quad \text{(meaning?)}$$

$$\text{Or}, CE(p,q) = H^p(X) + KL(p||q)$$



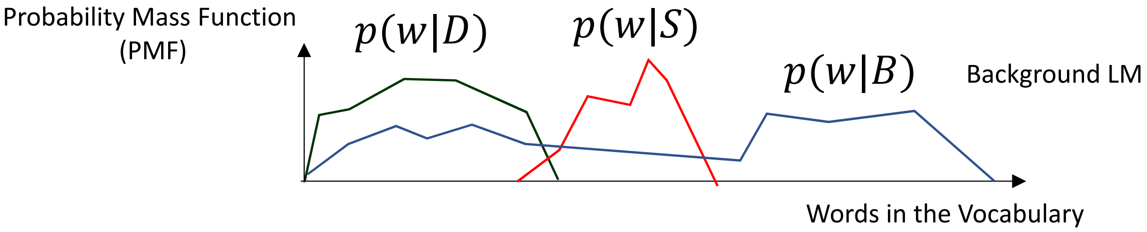
Document Summarization with KL Divergence, Cross Entropy and Entropy (1/2)

- We can use KL divergence to quantify how close a document D and one of its sentences S are
  - The closer the sentence model $p(w|S)$ to the document model $p(w|D)$, the more likely the sentence would be selected into the summary set

$$KL(D||S) = \sum_{w \in V} P(w|M_D) \log \frac{P(w|M_D)}{P(w|M_S)}$$

The lower the *KL* score, the more important S is!

  - A sentence *S* has a smaller value in terms of $KL(D||S)$ is deemed to be more important

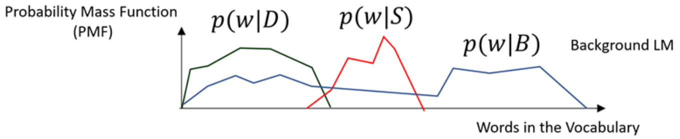


S.-H. Lin et al., “Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization,” IEEE Transactions on Audio, Speech and Language Processing, 2011

Document Summarization with KL Divergence, Cross Entropy and Entropy (2/2)

- Further, we can quantify the thematic specificity of each candidate summary sentence S, which is formally defined as follows

$$Clarity(S) \overset{\text{def}}{=} CE(B,S) - H(S)$$



  - Where B designates the background document collection
  - It is hypothesized that the higher the cross entropy (or the farther S away from the B), the more thematic information S is to convey
  - The lower the entropy $H(S)$, the more concentrative the word usage of the sentence S (S assigns higher probabilities to only some specific content words)
- The original KL divergence can be used in conjunction with the sentence-level clarity measure for important sentence ranking

$$Score(S) = -KL(D||S) + Clarity(S)$$

The higher the score, the more important S is !

S.-H. Liu et al., "Combining relevance language modeling and clarity measure for extractive speech summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015

LM for Information Retrieval (IR): Minimum KL Divergence

- Documents are ranked by Kullback-Leibler (KL) divergence (in increasing order)

$$KL(Q\|D) = \sum_{w} P(w|Q) \log \frac{P(w|Q)}{P(w|D)}$$

Query model

Document model

$$= \boxed{\sum_{w} P(w|Q) \log P\,(w|Q)} - \boxed{\sum_{w} P(w|Q) \log P\,(w|D)}$$

The same for all document
=> can be disregarded

Cross entropy between the language models of a query and a document

Equivalent to ranking **in decreasing order** of

$$\sum_{w} P(w|Q) \log P\,(w|D)$$

Relevant documents are deemed to have lower cross entropies

$$\overset{\text{rank}}{=} \sum_{w} c(w,Q) \log P\,(w|D) \overset{\text{rank}}{=} \log P(Q|D) \overset{\text{rank}}{=} P(Q|D)$$

For IR, minimum KL Divergence is equivalent to minimum cross entropy and maximum likelihood

Human Languages: Entropy Rate

- We can assume that a language L is a stochastic process consisting of a sequence of tokens $L = (X_1, X_2, \dots, X_n)$, then the entropy rate of L is

$$H_{\text{rate}}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

  - We take the entropy rate of a language to be the limit of the entropy rate of a sample of the language as the sample gets longer and longer

Human Languages: Cross Entropy Rate

- We can also define the cross entropy rate of a language $L = (X) \sim p(x_{1n})$ according to a model m by

$$CE_{\text{rate}}(L, m) = -\lim_{n\to\infty} \frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log m\,(x_{1n})$$

$$\approx -\lim_{n\to\infty} \frac{1}{n} \log m(x_{1n}) \qquad \text{((when } n \to \infty, p(x_{1n})\text{=1 ?))}$$

$$\approx -\frac{1}{n} \log m(x_{1n}) \qquad \text{(when n is large enough)}$$

$$= -\frac{1}{n} \sum_{j=1}^{n} \log m(x_j | x_1, x_2, \dots, x_{j-1})$$

$$= \log \sqrt[n]{\prod_{j=1}^{n} \frac{1}{m(x_j | x_1, x_2, \dots, x_{j-1})}}$$

Note that: $\text{Perplexity}(x_{1n}, m) = 2^{CE_{\text{rate}}(x_{1n}, m)} = m(x_{1n})^{-\frac{1}{n}}$