Foundations

Computer Science- Pattern Recognition

Prof. Dr. Dhahir A. Abdullah

OUTLINE

- Probability Theory
- Linear Algebra

Probability Theory: Sets

- Probability makes extensive use of set operations,
- A **set** is a collection of objects, which are the **elements** of the set,
- If S is a set and x is an element of S, we write $x \in S$.
- If x is not an element of S, we write $x /\in S$.
- A set can have no elements, in which case it is called the **empty set**, denoted by Ø.

Probability Theory: Sets

- Sets can be specified as

$$S = \{x_1, x_2, \ldots, x_n\}.$$

- For example,
  - the set of possible outcomes of a die roll is *{1, 2, 3, 4, 5, 6}*,
  - The set of possible outcomes of a coin toss is *{H, T}*, where *H* stands for "heads" and *T* stands for "tails."

Probability Theory: Sets

- Alternatively, we can consider the set of all *x* that have a certain property *P*, and denote it by

$$\{x \,|\, x \text{ satisfies } P\}.$$

- (The symbol “/” is to be read as “such that.”)
- For example,
  - the set of even integers can be written as *{k / k/*2 is integer*}*.

Probability Theory: Sets Operations

- **Complement:**
  - The **complement** of a set *S*, with respect to the universe Ω, is the set $\{x \in \Omega / x /\in S\}$ of all elements of Ω that do not belong to *S*, and is denoted by S^c.
- **Union:**

$$S \cup T = \{x \,|\, x \in S \text{ or } x \in T\},$$

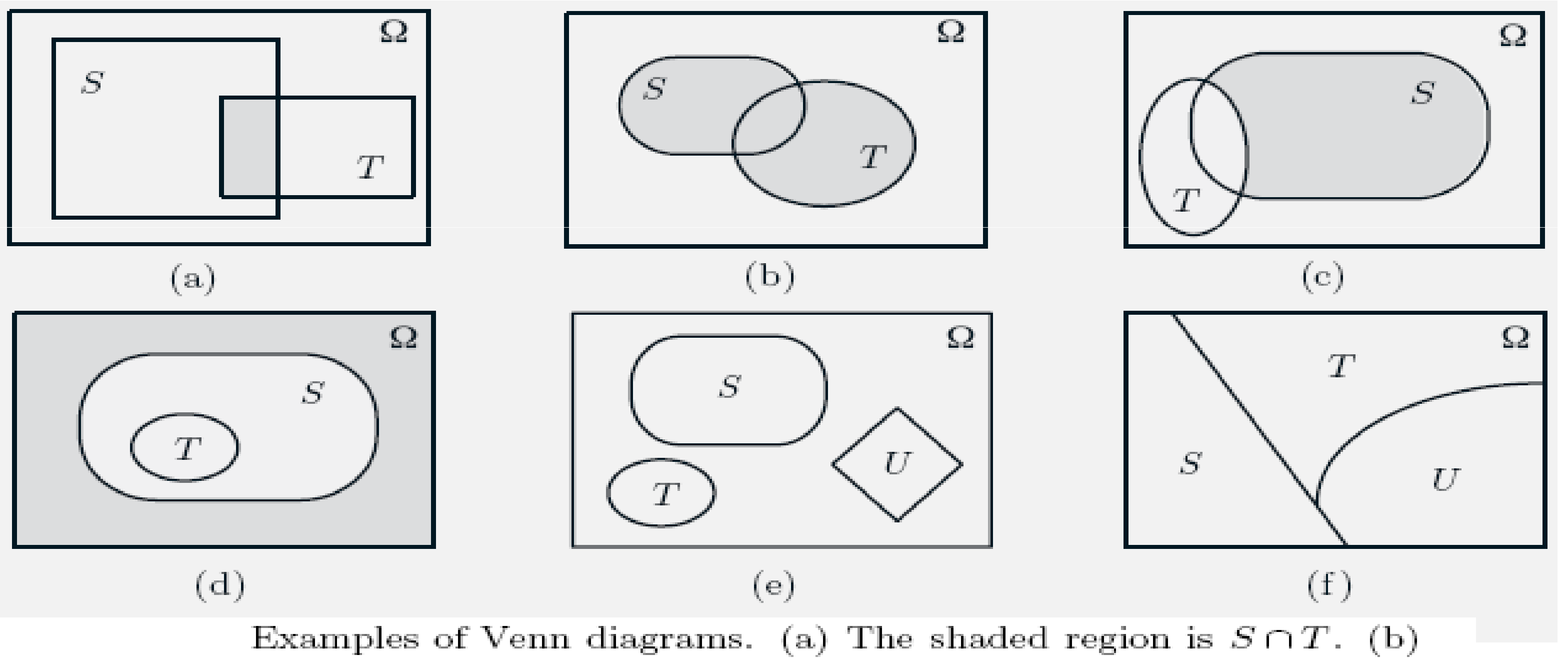
- **Intersection:**

$$S \cap T = \{x \,|\, x \in S \text{ and } x \in T\}.$$

Probability Theory: Sets Operations

- **Disjoint:**
  - several sets are said to be **disjoint** if no two of them have a common element
- **Partition:**
  - A collection of sets is said to be a **partition** of a set *S* if the sets in the collection are disjoint and their union is *S*.

Probability Theory: Sets Operations



Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of the set Ω.

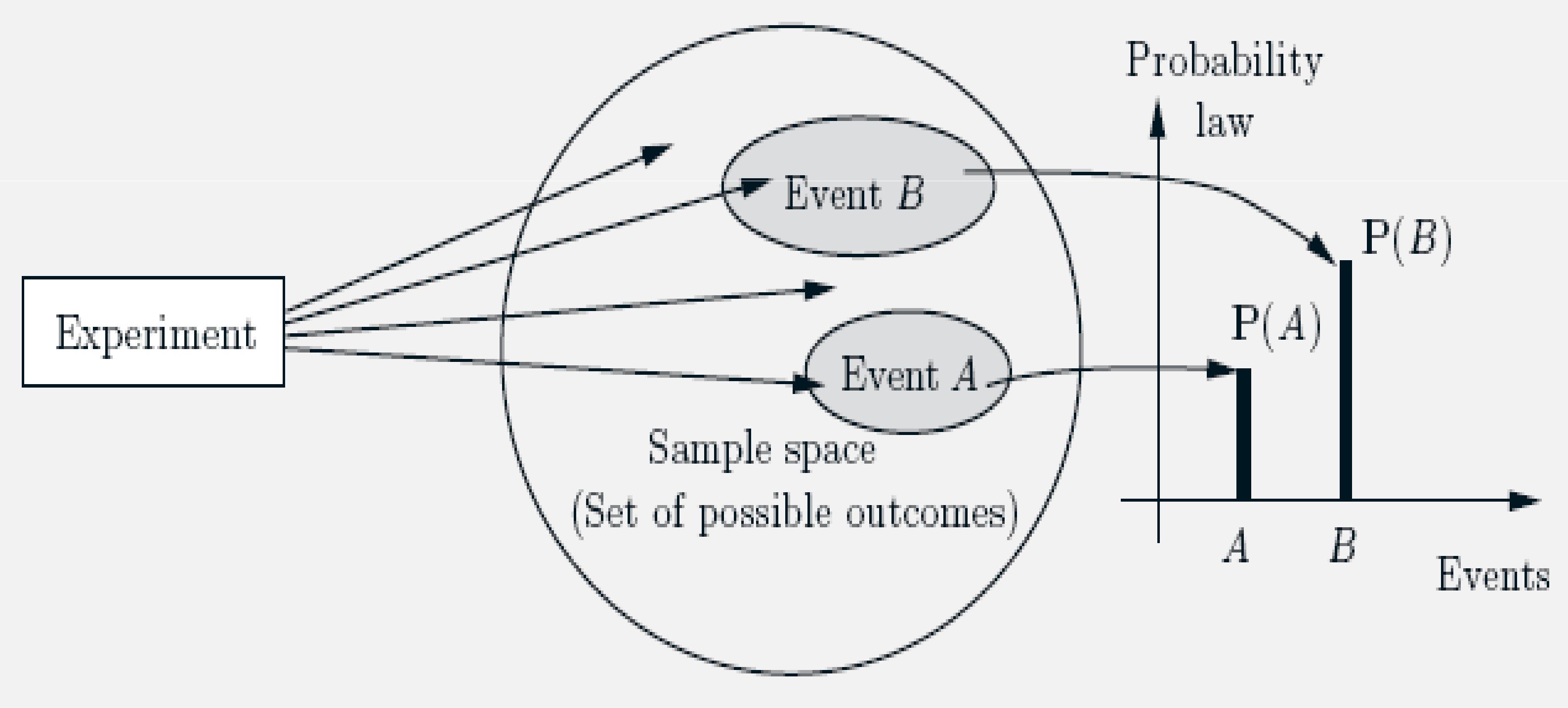
Probability Theory

- The set of all possible outcomes of an experiment is the **sample space**, denoted **Ω**.
- An **event A** is a (set of) possible outcomes of the experiment, and corresponds to a subset of **Ω**.

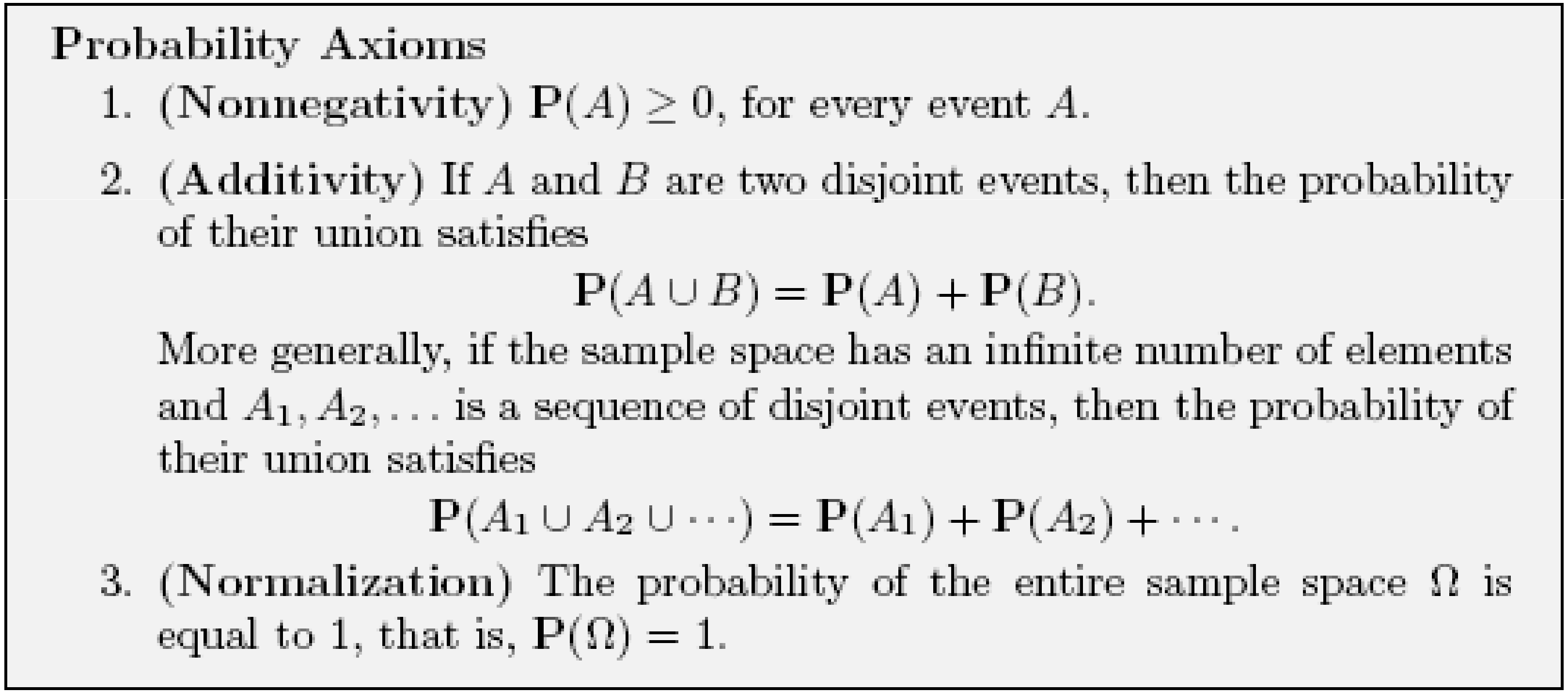
Probability Theory

- A probability law / measure is a function **P(A)**
  - with the argument **A,**
  - that assigns a value to **A** based on the expected proportion of number of times that event **A** is actually likely to happen.

Probability Theory



Probability Theory: Axioms of Probability



Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A.

2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

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

   More generally, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots.$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

Probability Theory: Axioms of Probability

- The probability function **P(A)** must satisfy the following:

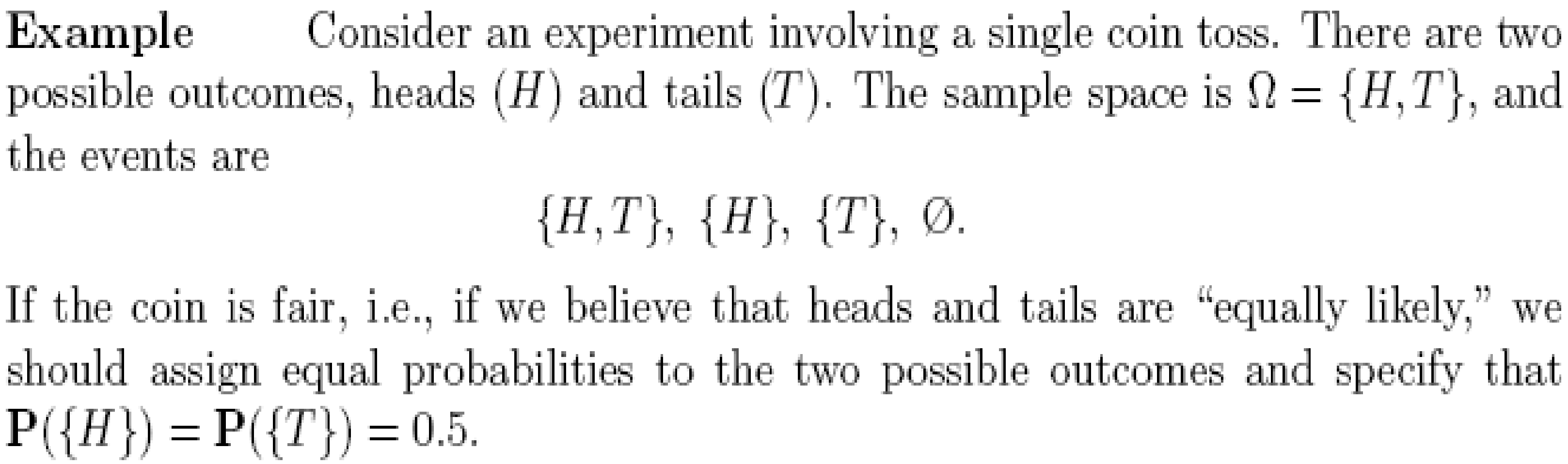
$$0 \le P(A_i) \le 1,$$

$$P(\Omega) = \sum P(A_i)$$

$$A_i \cap A_j = \phi \Rightarrow \quad P(A_i \cup A_j) = P(A_i) + P(A_j),$$

$$A_i \cap A_j \ne \phi \Rightarrow \quad P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j),$$

Probability Theory: Axioms of Probability (Example)

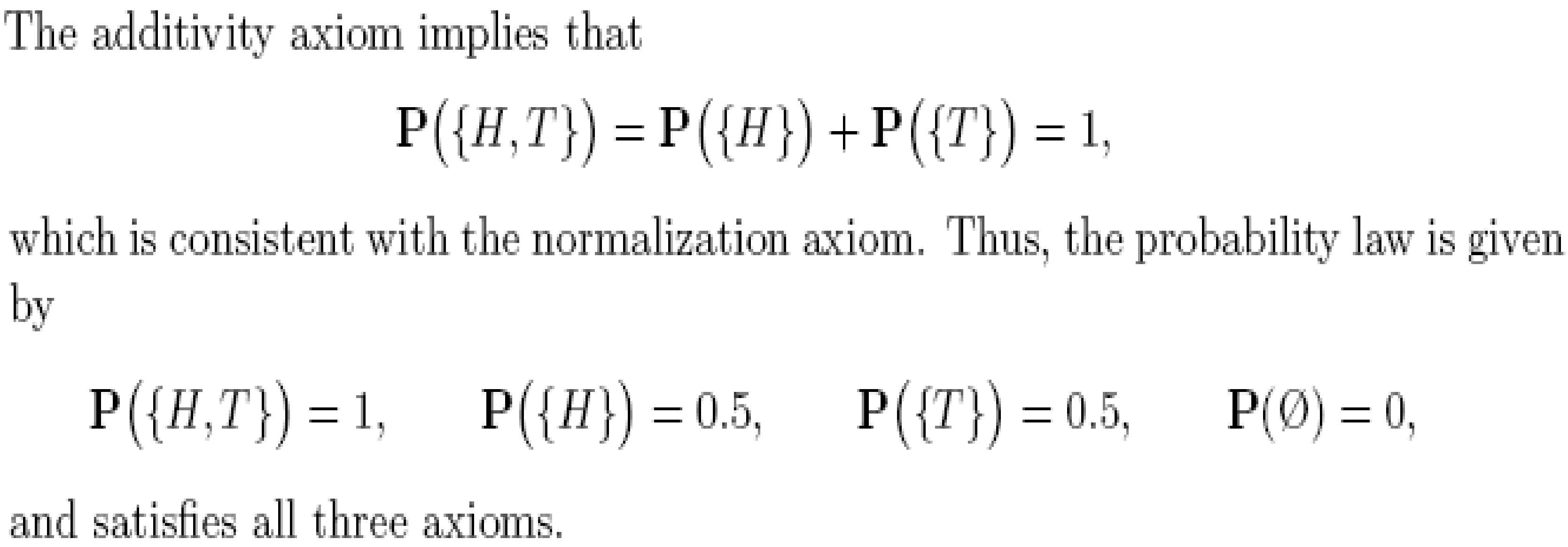


Example Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T). The sample space is $\Omega = \{H, T\}$, and the events are

$$\{H, T\},\ \{H\},\ \{T\},\ \emptyset.$$

If the coin is fair, i.e., if we believe that heads and tails are "equally likely," we should assign equal probabilities to the two possible outcomes and specify that $\mathbf{P}(\{H\}) = \mathbf{P}(\{T\}) = 0.5$.

Probability Theory: Axioms of Probability (Example)



The additivity axiom implies that

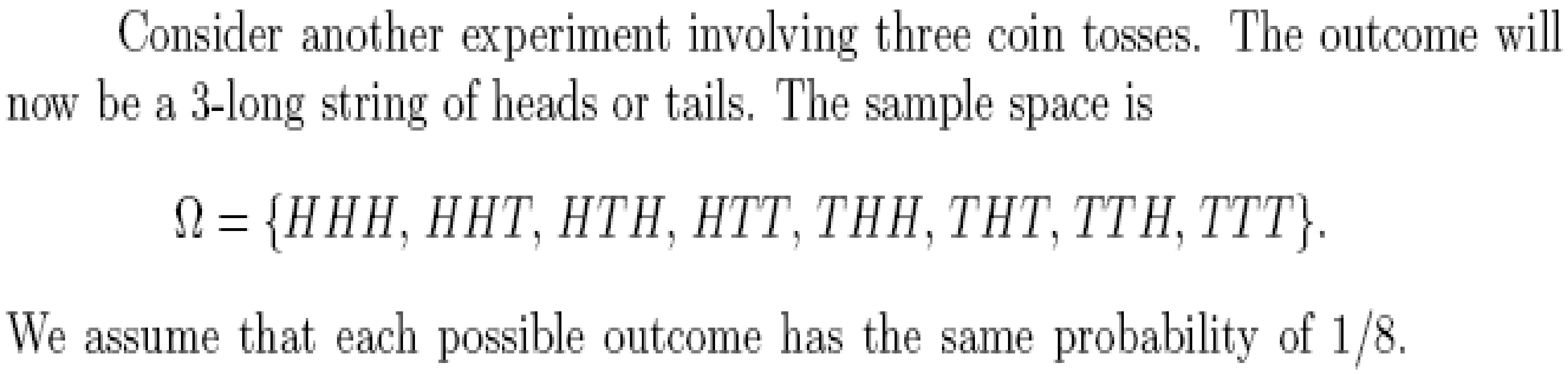
$$\mathbf{P}\big(\{H,T\}\big) = \mathbf{P}\big(\{H\}\big) + \mathbf{P}\big(\{T\}\big) = 1,$$

which is consistent with the normalization axiom. Thus, the probability law is given by

$$\mathbf{P}\big(\{H,T\}\big) = 1, \qquad \mathbf{P}\big(\{H\}\big) = 0.5, \qquad \mathbf{P}\big(\{T\}\big) = 0.5, \qquad \mathbf{P}(\varnothing) = 0,$$

and satisfies all three axioms.

Probability Theory: Axioms of Probability (Example)

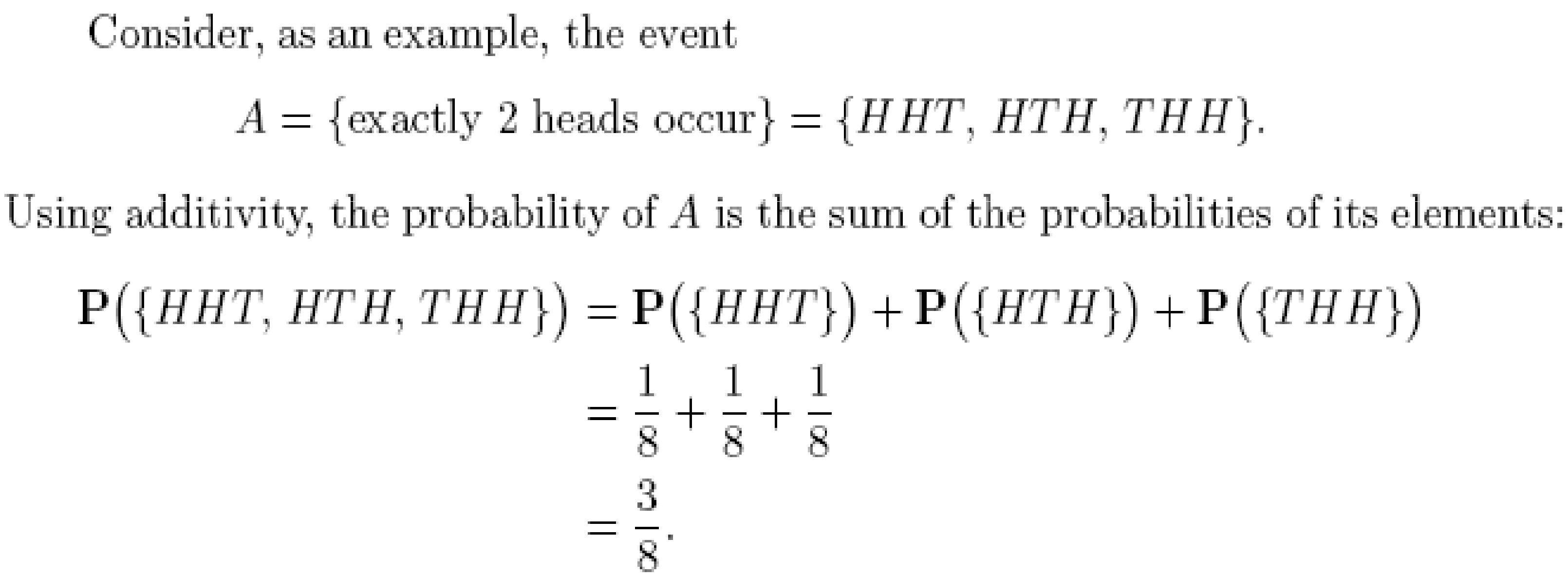


Consider another experiment involving three coin tosses. The outcome will now be a 3-long string of heads or tails. The sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

We assume that each possible outcome has the same probability of $1/8$.

Probability Theory: Axioms of Probability (Example)



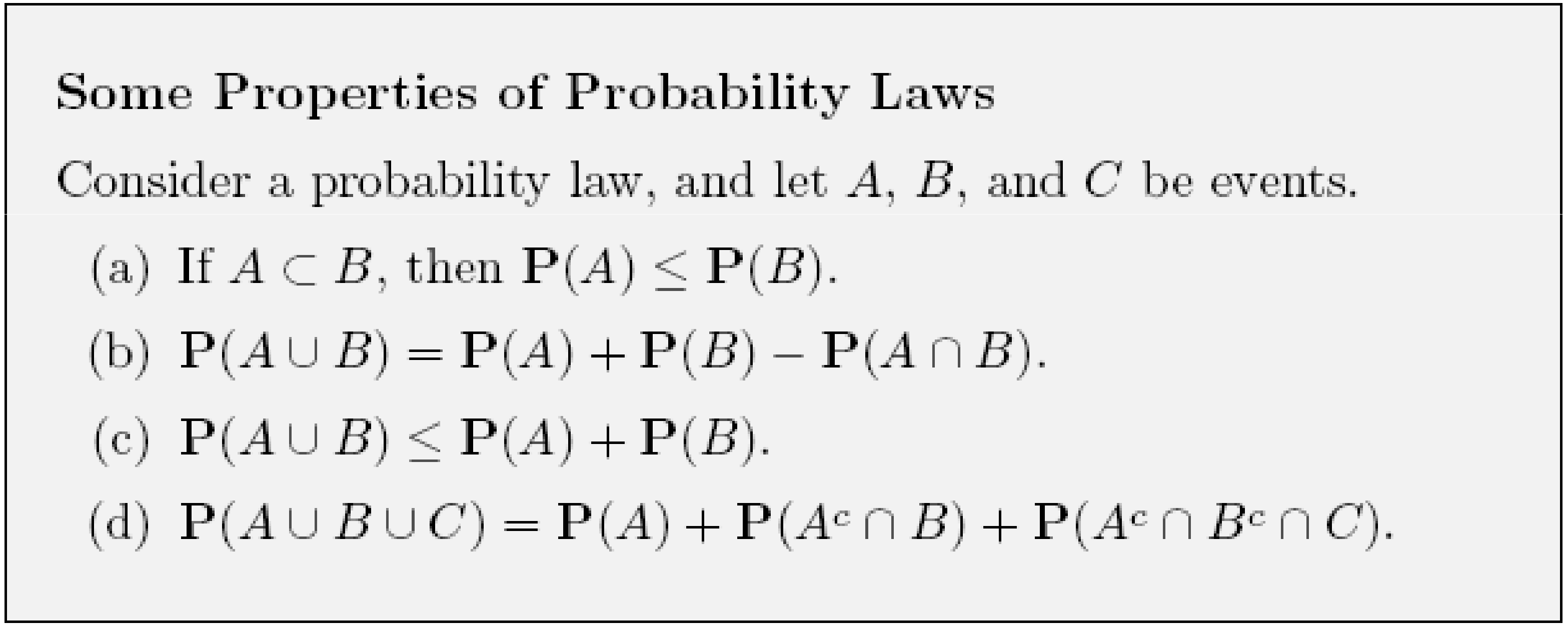
Consider, as an example, the event

$$A = \{\text{exactly 2 heads occur}\} = \{HHT, HTH, THH\}.$$

Using additivity, the probability of A is the sum of the probabilities of its elements:

$$\begin{aligned}
\mathbf{P}\big(\{HHT, HTH, THH\}\big) &= \mathbf{P}\big(\{HHT\}\big) + \mathbf{P}\big(\{HTH\}\big) + \mathbf{P}\big(\{THH\}\big) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\
&= \frac{3}{8}.
\end{aligned}$$

Probability Theory



Some Properties of Probability Laws

Consider a probability law, and let A, B, and C be events.

(a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

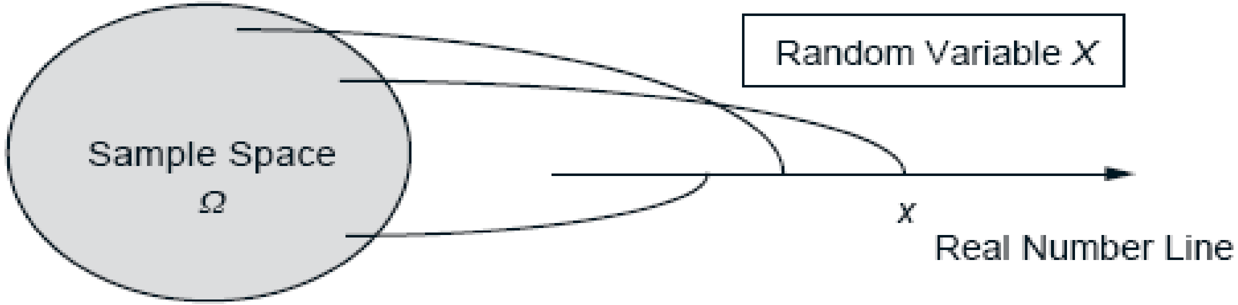
(d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

Probability Theory: Random Variables

- In many probabilistic models, the outcomes are of a numerical nature, e.g., if they correspond to instrument readings or stock prices. In other experiments,
- the outcomes are not numerical, but they may be associated with some numerical values of interest. For example, if the experiment is the selection of students from a given population, we may wish to consider their grade point average.
- When dealing with such numerical values, it is often useful to assign probabilities to them.
- This is done through the notion of a **random variable.**

Probability Theory: Random Variables

- Briefly:
  - A random variable X is a function that maps every possible event in the space Ω of a random experiment to a real number.



Probability Theory: Random Variables

- Random variables can **discrete**, e.g., the number of heads in three consecutive coin tosses, or **continuous**, the weight of a class member.

Probability Theory: Random Variables

- Random variables can
  - **discrete**, e.g., the number of heads in three consecutive coin tosses, or
  - **continuous**, the weight of a class member.

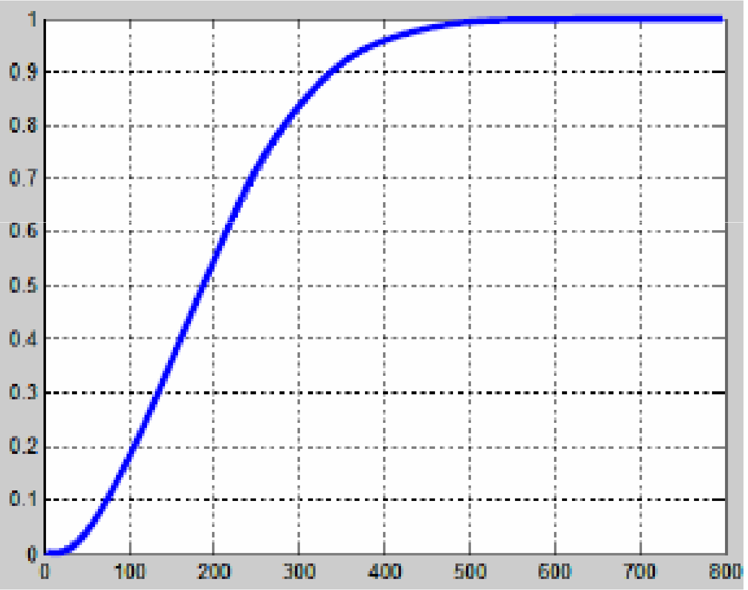
Probability Theory: Probability/Cumulative Mass Function

- A **probability mass (distribution) function** is a function that tells us the probability of x, an observation of X, assuming a specific value.
  - $P(X=x)$
- The **cumulative mass (distribution) function** indicates the probability of X assuming a value less then or equal to x.
  - $F(x) = P(X<=x)$

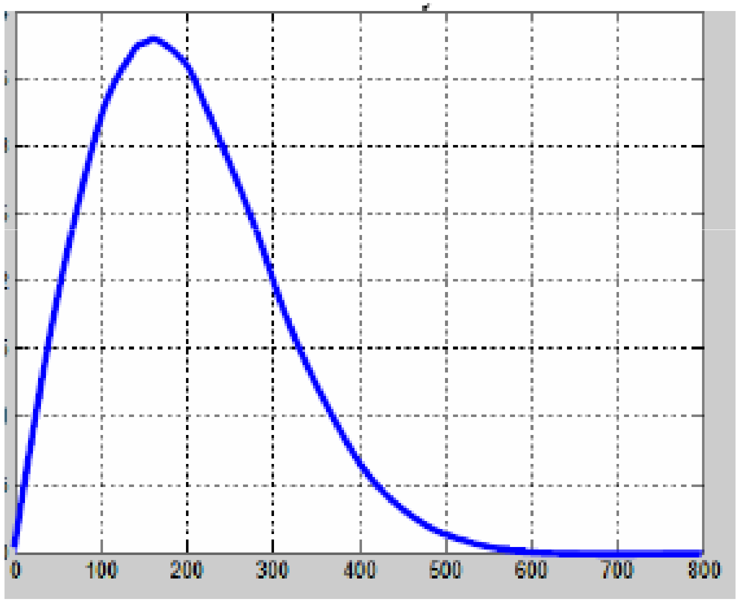
Probability Theory: Probability/Cumulative Density Function

- For continuos variables:
  - Probability Mass Function → Probability Density function,
  - Cumulative Mass Function → Cumulative Density Function

Probability Theory: Probability/Cumulative Density Function



cdf



pdf

Expectation

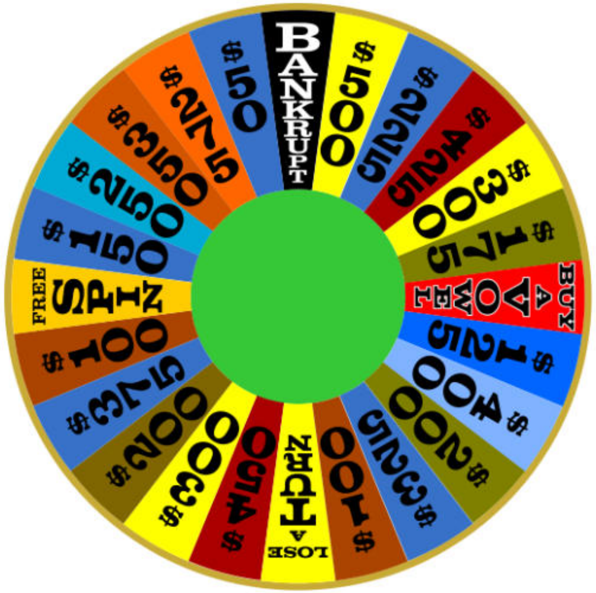
- The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X. It would be desirable to summarize this information in a single representative number.
- This is accomplished by the **expectation** of X, which is a weighted (in proportion to probabilities) average of the possible values of X.

Expectation

- Example:
  - suppose you spin a wheel of fortune many times,
  - at each spin, one of the numbers $m1,m2, . . . , mn$ comes up with corresponding probability $p1, p2, . . . , pn$, and
  - this is your reward from that spin.

Expectation

- Example:
  - What is the amount of money that you “expect” to get “per spin”?



Expectation

- Example:
  - Suppose that you spin the wheel ***k* times**, and that ***ki*** is the number of times that the outcome is ***mi***. Then, the total amount received is ***m*1*k*1 +*m*2*k*2 +· · · + *mnkn***. The amount received per spin is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k}.$$

Expectation

- Example:
  - If the number of spins ***k*** **is very large**, and if we are willing to interpret probabilities as **relative frequencies**, it is reasonable to anticipate that ***mi*** comes up a fraction of times that is roughly equal to ***pi***:

$$p_i \approx \frac{k_i}{k}, \qquad i = 1, \ldots, n.$$

Expectation

- Example:
  - Thus, the **amount of money per spin** that you “expect” to receive is

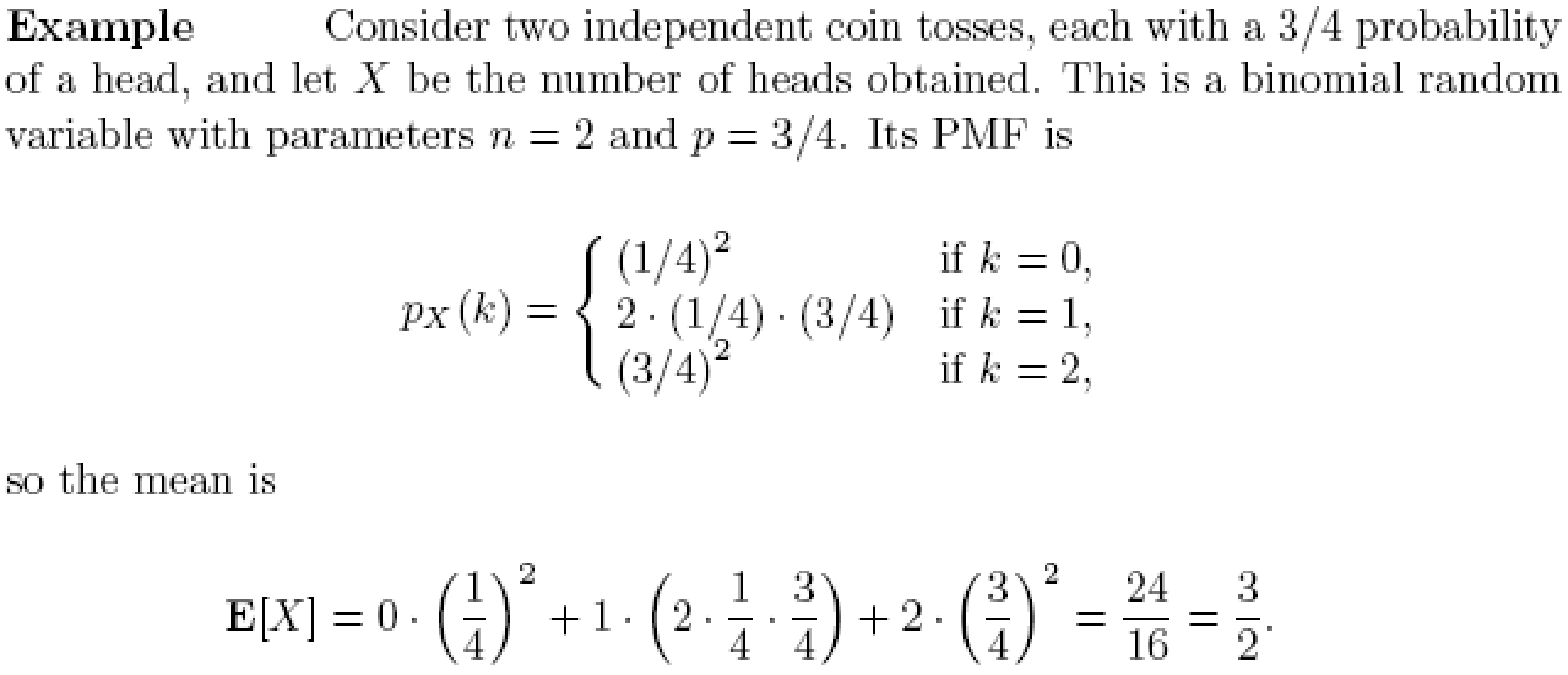
$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k}$$
$$\approx m_1p_1 + m_2p_2 + \cdots + m_np_n.$$

Expectation

- The expected value, or **average**, of a random variable **X**, whose possible values are **{x1,…,xm}** with respective probabilities **p1,…,pm**, is given as:

$$\boxed{E(x) = \mu = \sum_{x \in X} x \cdot P(x) = \sum_{i=1}^{m} x_i \cdot p_i}$$

Expectation



Example Consider two independent coin tosses, each with a 3/4 probability of a head, and let X be the number of heads obtained. This is a binomial random variable with parameters $n = 2$ and $p = 3/4$. Its PMF is

$$p_X(k) = \begin{cases} (1/4)^2 & \text{if } k = 0, \\ 2 \cdot (1/4) \cdot (3/4) & \text{if } k = 1, \\ (3/4)^2 & \text{if } k = 2, \end{cases}$$

so the mean is

$$\mathbf{E}[X] = 0 \cdot \left(\frac{1}{4}\right)^2 + 1 \cdot \left(2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) + 2 \cdot \left(\frac{3}{4}\right)^2 = \frac{24}{16} = \frac{3}{2}.$$

Expectation

- Moments of a random variable:

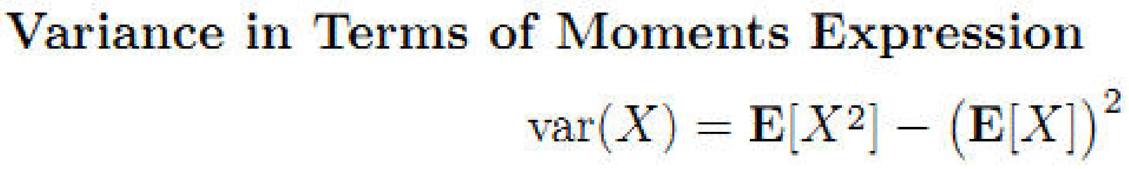
$$\boxed{E(x^k) = \sum_{x \in X} x^k \cdot P(x) = \sum_{i=1}^{m} (x_i)^k \cdot p_i}$$

Expectation

- the **variance**, the average dispersion of the data from the mean

$$Var[x] = \sigma^2 = E\left((x-\mu)^2\right) = \sum_{x\in X} (x-\mu)^2 \cdot P(x)$$

Expectation

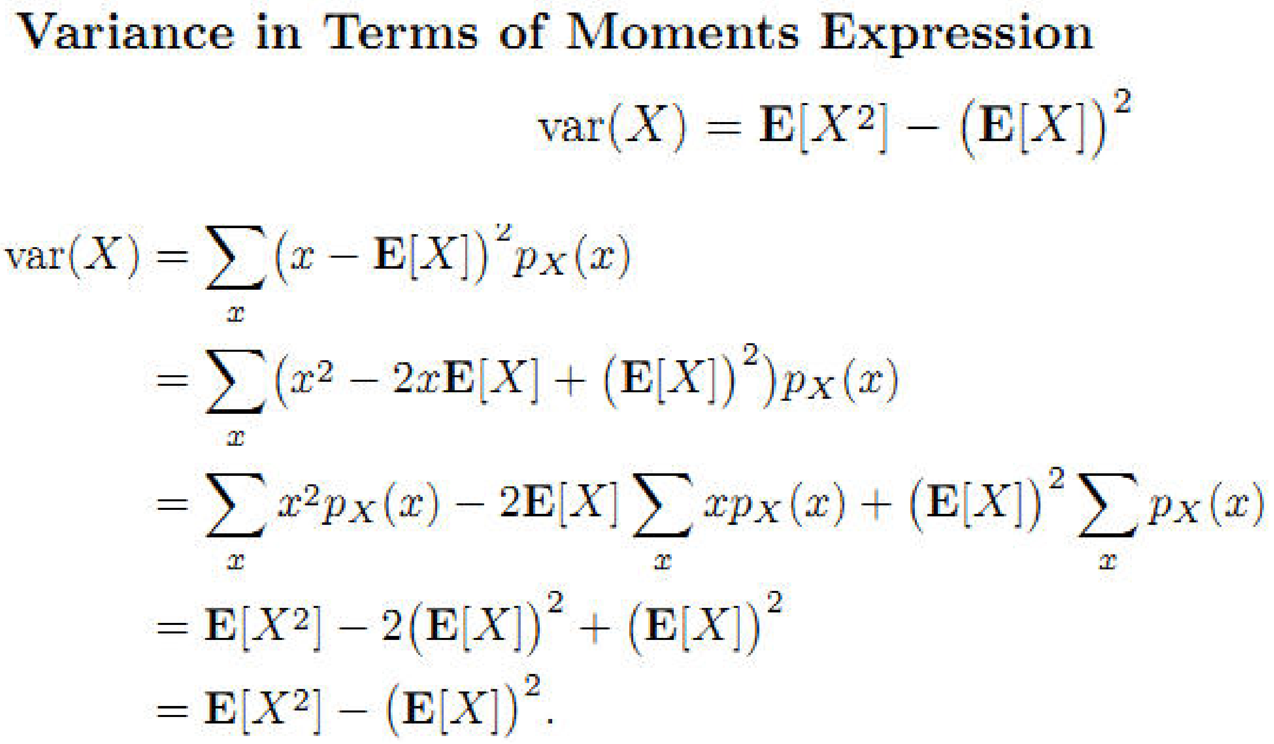


Variance in Terms of Moments Expression

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \left(\mathbf{E}[X]\right)^2$$

How ?

Expectation

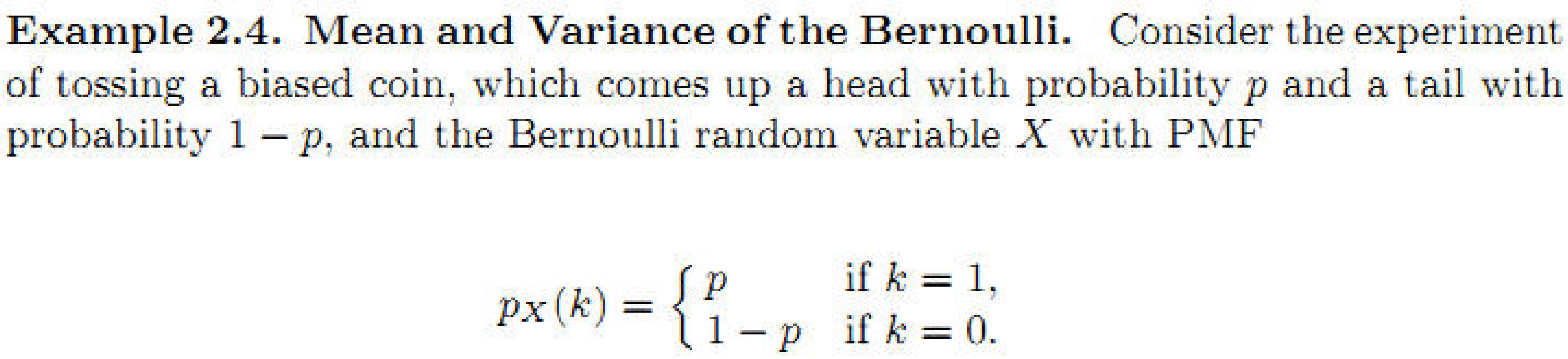


Variance in Terms of Moments Expression

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \left(\mathbf{E}[X]\right)^2$$

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

Expectation

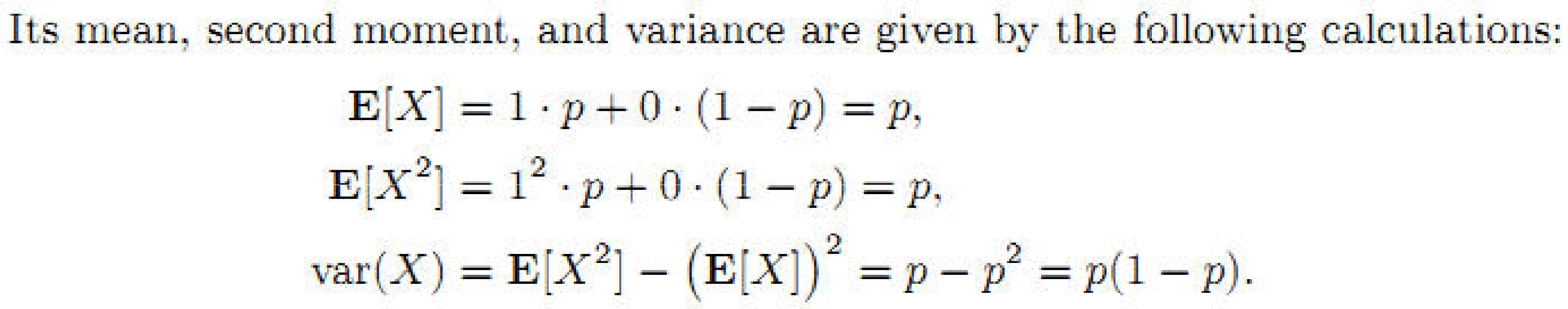


Example 2.4. Mean and Variance of the Bernoulli. Consider the experiment of tossing a biased coin, which comes up a head with probability p and a tail with probability $1 - p$, and the Bernoulli random variable X with PMF

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

Mean, second moment, variance ?

Expectation



Its mean, second moment, and variance are given by the following calculations:

$$\begin{aligned}\mathbf{E}[X] &= 1 \cdot p + 0 \cdot (1-p) = p,\\ \mathbf{E}[X^2] &= 1^2 \cdot p + 0 \cdot (1-p) = p,\\ \mathrm{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = p - p^2 = p(1-p).\end{aligned}$$

Expectation

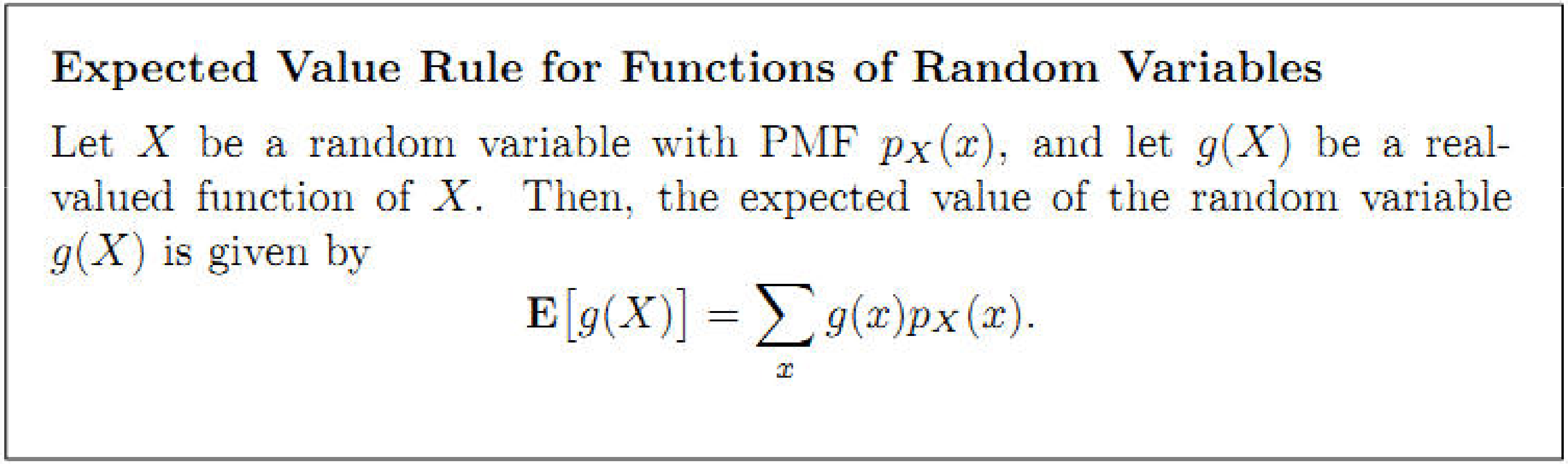
- The variance provides a measure of dispersion of X around its mean. An-other measure of dispersion is the **standard deviation** of X, which is defined as the square root of the variance

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

Expectation

- **The standard deviation is often easier to interpret**, because it has the same units as X.
- For example, if X measures length in meters, the units of variance are square meters, while the units of the standard deviation are meters.

Expectation



Expected Value Rule for Functions of Random Variables

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

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

Pairs of Random Variables

- Consider two discrete random variables X and Y associated with the same experiment. The joint PMF of X and Y is defined by

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

Pairs of Random Variables

- Joint probability also need to satisfy the axioms of the probability theory

$$P(x,y) > 0 \quad and \quad \sum_{x\in X}\sum_{y\in Y} P(x,y) = 1$$

- Everything that relate to X, or Y – individually or together – can be obtained from the P(x,y). In particular, the individual pmfs, called the marginal distribution functions can be obtained as

$$P_x(x) = \sum_{y\in Y} P(x,y) \qquad P_y(y) = \sum_{x\in X} P(x,y)$$

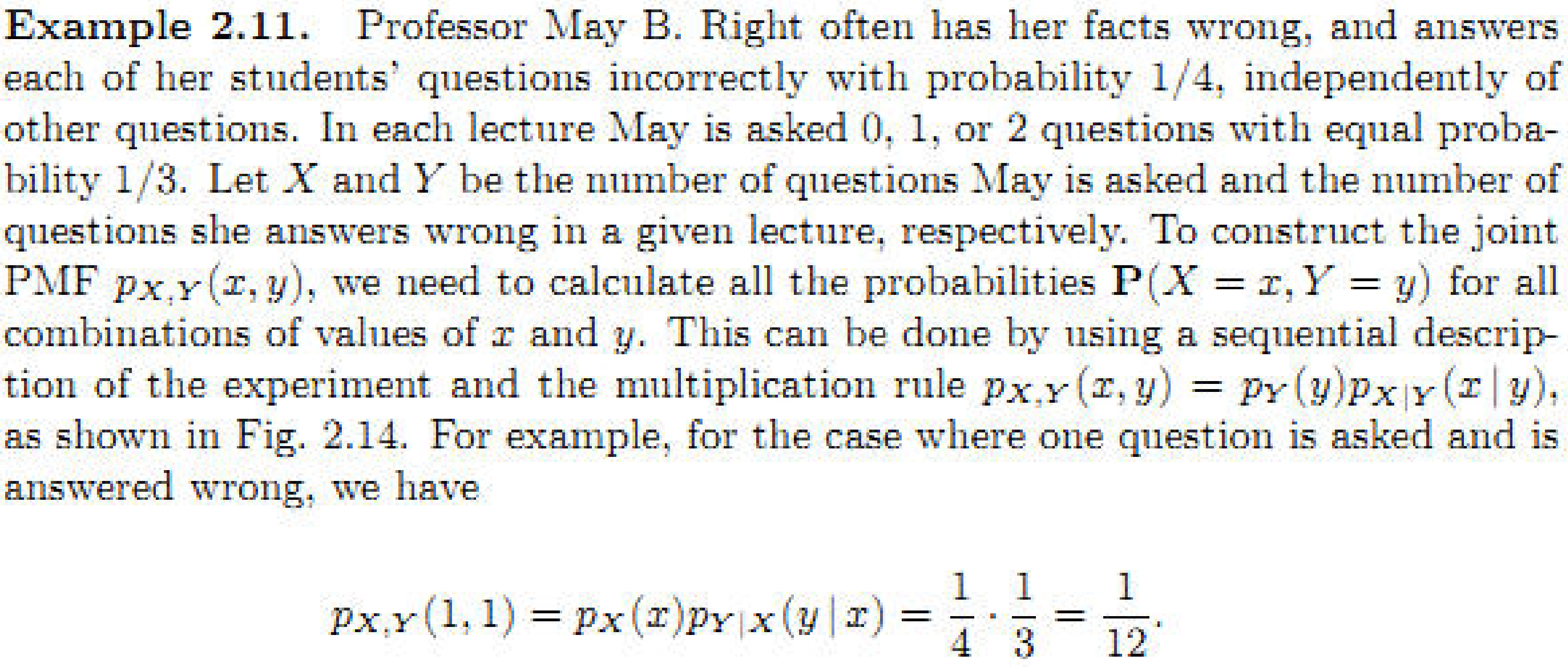
Statistical Independence

- **Random variables X and Y are said to be statistically independent, if and only if**

$$\boxed{P(x, y) = P_x(x) \cdot P_y(y)}$$

- That is, if the outcome of one event does not effect the outcome of the other, they are statistically independent. For example, the outcome of two individual dice are independent, as one does not affect the other.

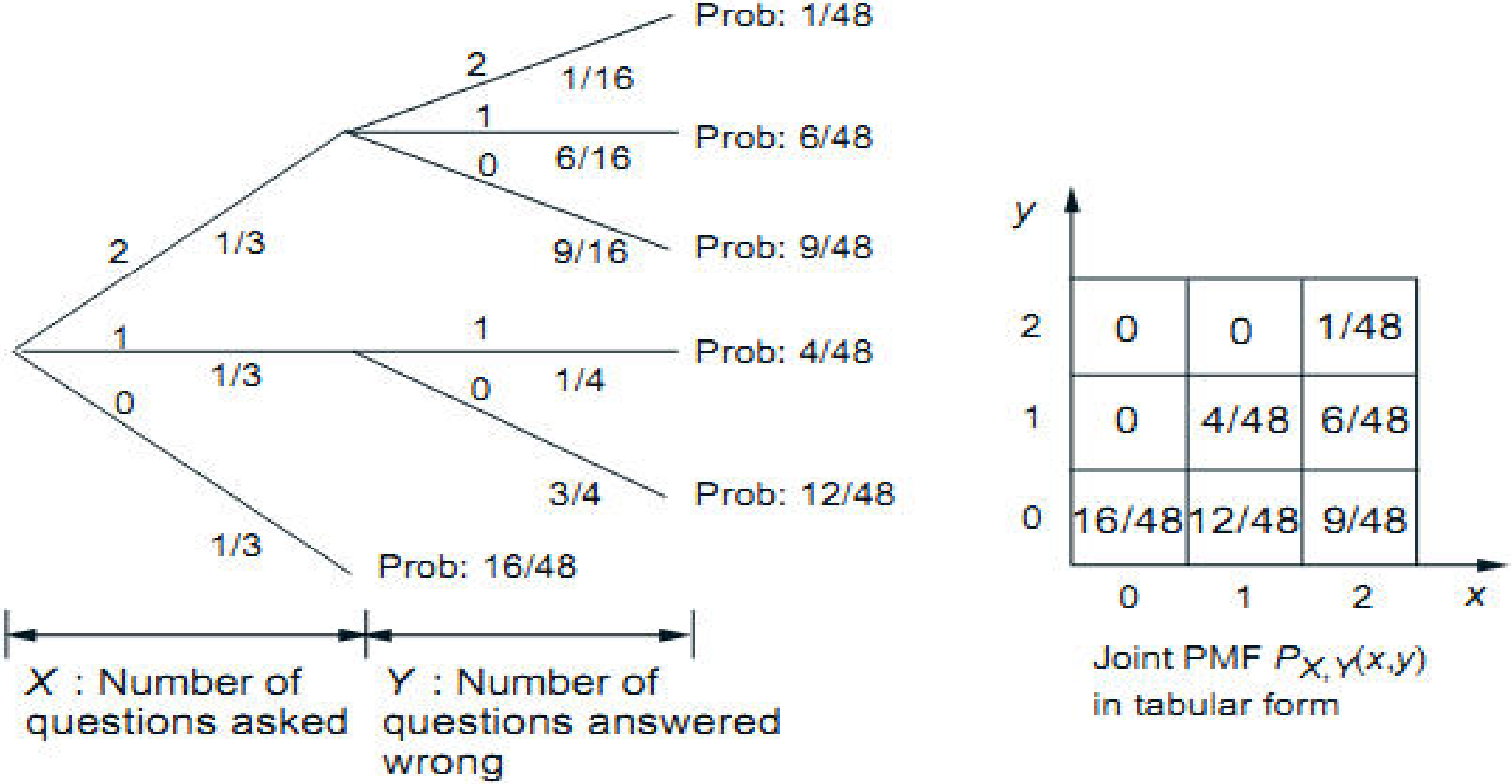
Example



Example 2.11. Professor May B. Right often has her facts wrong, and answers each of her students' questions incorrectly with probability 1/4, independently of other questions. In each lecture May is asked 0, 1, or 2 questions with equal probability 1/3. Let X and Y be the number of questions May is asked and the number of questions she answers wrong in a given lecture, respectively. To construct the joint PMF $p_{X,Y}(x,y)$, we need to calculate all the probabilities $\mathbf{P}(X=x, Y=y)$ for all combinations of values of x and y. This can be done by using a sequential description of the experiment and the multiplication rule $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x\,|\,y)$, as shown in Fig. 2.14. For example, for the case where one question is asked and is answered wrong, we have

$$p_{X,Y}(1,1) = p_X(x)p_{Y|X}(y\,|\,x) = \frac{1}{4}\cdot\frac{1}{3} = \frac{1}{12}.$$

Example (cont.)



y \ x	0	1	2
2	0	0	1/48
1	0	4/48	6/48
0	16/48	12/48	9/48

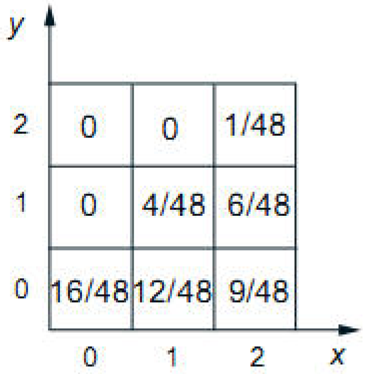
Joint PMF $P_{X,Y}(x,y)$ in tabular form

Figure 2.14: Calculation of the joint PMF $p_{X,Y}(x, y)$ in Example 2.11.

Example (cont.)

The joint PMF can be represented by a two-dimensional table, as shown in Fig. 2.14. It can be used to calculate the probability of any event of interest. For instance, we have

$$\begin{aligned}\mathbf{P}(\text{at least one wrong answer}) &= p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(2,2)\\ &= \frac{4}{48} + \frac{6}{48} + \frac{1}{48}.\end{aligned}$$



y \ x	0	1	2
2	0	0	1/48
1	0	4/48	6/48
0	16/48	12/48	9/48

Joint PMF $P_{X,Y}(x,y)$
in tabular form

Pairs of Random Variables

- Expected values, moments and variances of joint distributions can be computed similar to single variable cases:

$$\mu_x = \sum_x \sum_y x \cdot P(x,y) \qquad \mu_y = \sum_x \sum_y y \cdot P(x,y)$$

$$\sigma_x^2 = E\left[(x-\mu_x)^2\right] = \sum_x \sum_y (x-\mu_x)^2 \cdot P(x,y)$$

$$\sigma_y^2 = E\left[(y-\mu_y)^2\right] = \sum_x \sum_y (y-\mu_y)^2 \cdot P(x,y)$$

Co-Variance

- A cross-moment can also be defined as the covariance

$$\sigma_{xy}^{2} = E[(x-\mu_x)(y-\mu_y)] = \sum_x \sum_y (x-\mu_x)\cdot(x-\mu_x)\cdot P(x,y)$$

- Covariance defines how the variables vary together as a pair – are they both increasing together, does one increase when the other decrease, etc

Co-Variance

- the covariance matrix, denoted by ∑

$$\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$$

Correlation Coefficient

$$\boxed{\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}} \qquad \boxed{-1 \le \rho \le 1}$$

- If **ρ=1**, then the variables are identical, they move together,
- If **ρ=-1**, then the variables are negatively correlated, one decreases as the other increases at the same rate
- If **ρ=0** the variables are uncorrelated. The variation of one, has no effect on the other.

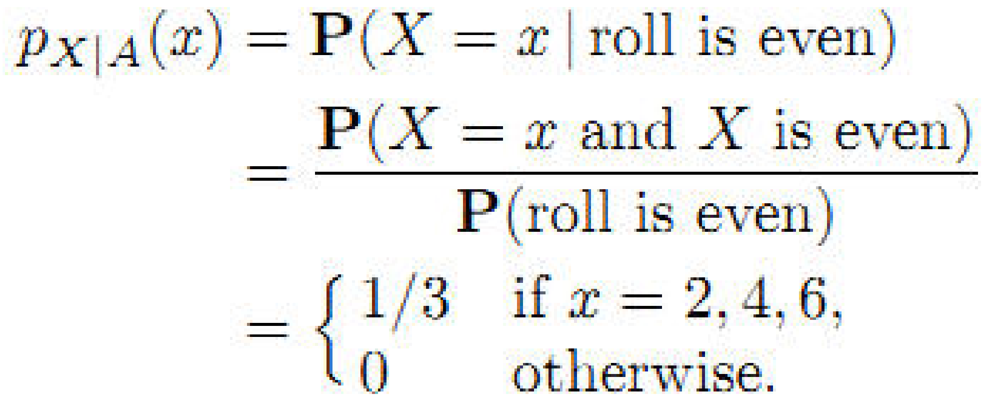
Conditional Probability

- The conditional PMF of a random variable X, conditioned on a particular event A with P(A) > 0, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x \,|\, A) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}.$$

Conditional Probability

- A is a legitimate PMF. As an example, let X be the **roll of a die** and let A be the event that the roll is an **even** number. Then, by applying the preceding formula, we obtain



$$\begin{aligned} p_{X|A}(x) &= \mathbf{P}(X = x \,|\, \text{roll is even}) \\ &= \frac{\mathbf{P}(X = x \text{ and } X \text{ is even})}{\mathbf{P}(\text{roll is even})} \\ &= \begin{cases} 1/3 & \text{if } x = 2, 4, 6, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Conditional Probability

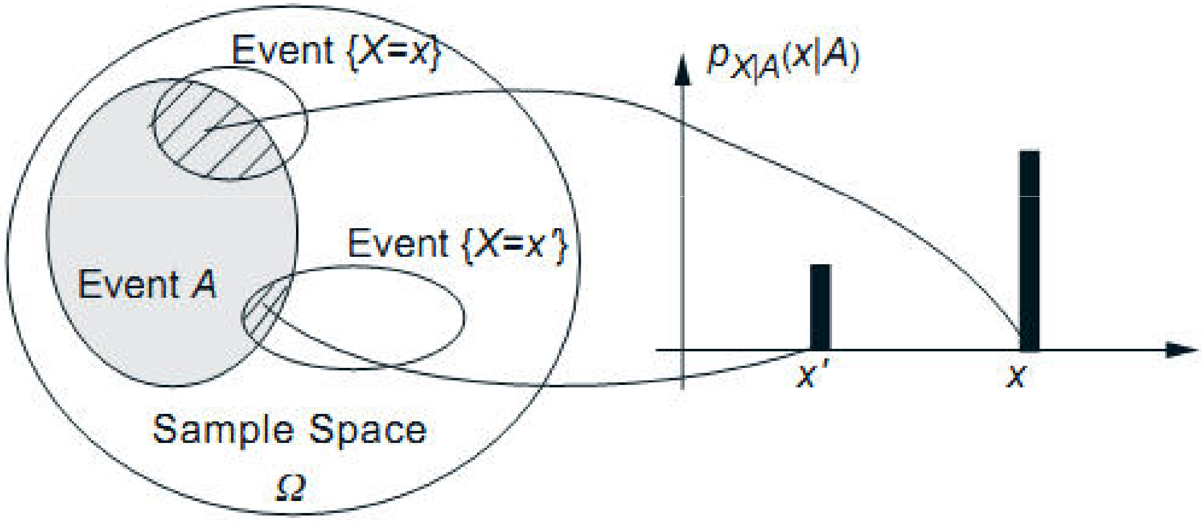
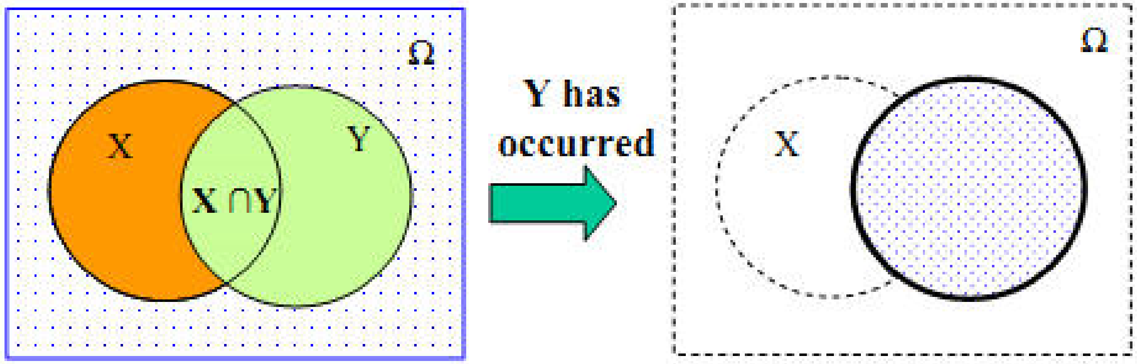


Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x, we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$ and normalize by diving with $\mathbf{P}(A)$.

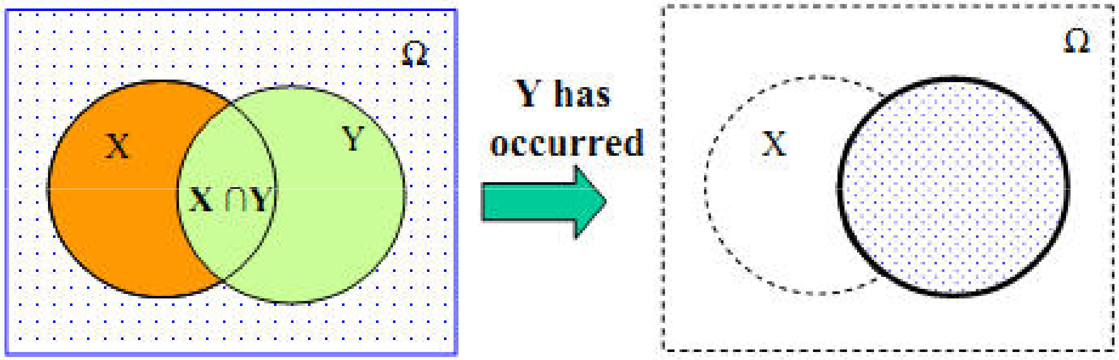
Conditional Probability

- The conditional probability of X=x given the Y=y has been observed is given as

$$\boxed{P(X=x\,|\,Y=y)=\frac{P(X=x\cap Y=y)}{P(Y=y)}=\frac{P(x,y)}{P(y)}}\;\Rightarrow\;\boxed{P(x,y)=P(x\,|\,y)P(y)}$$

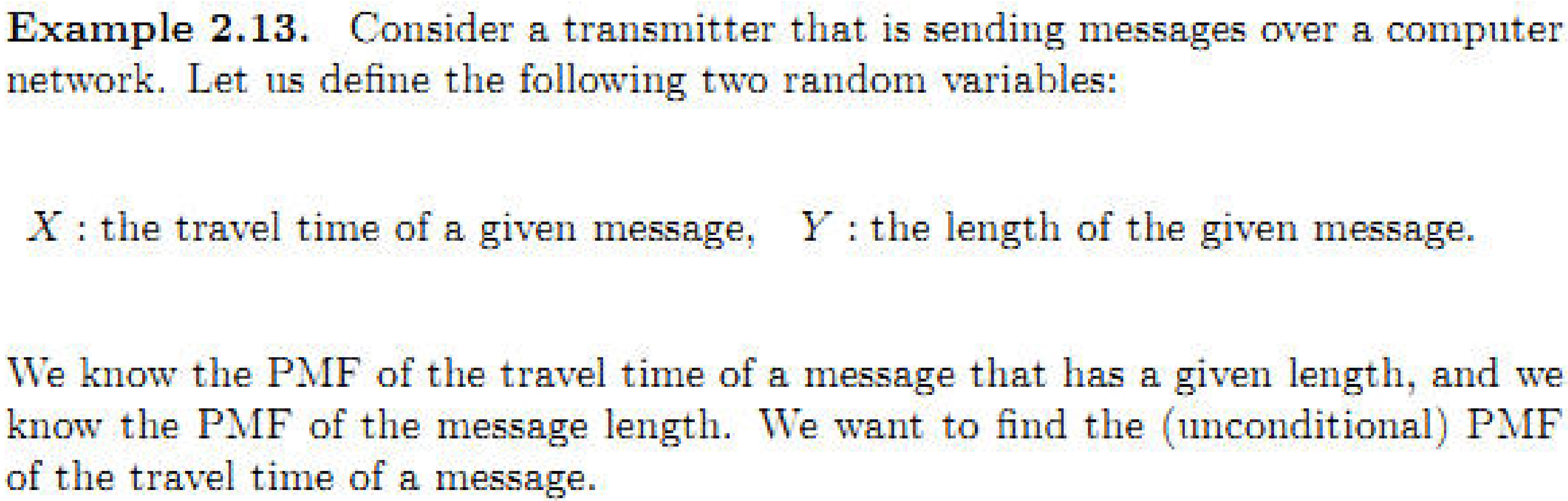


Conditional Probability



- The fact that Y=y has been observed has two main consequences:
  - The sample space effectively becomes the space of Y
  - The event X=x, effectively becomes X∩Y, that is P(y) renormalizes the probability of events that occur jointly with Y

Example

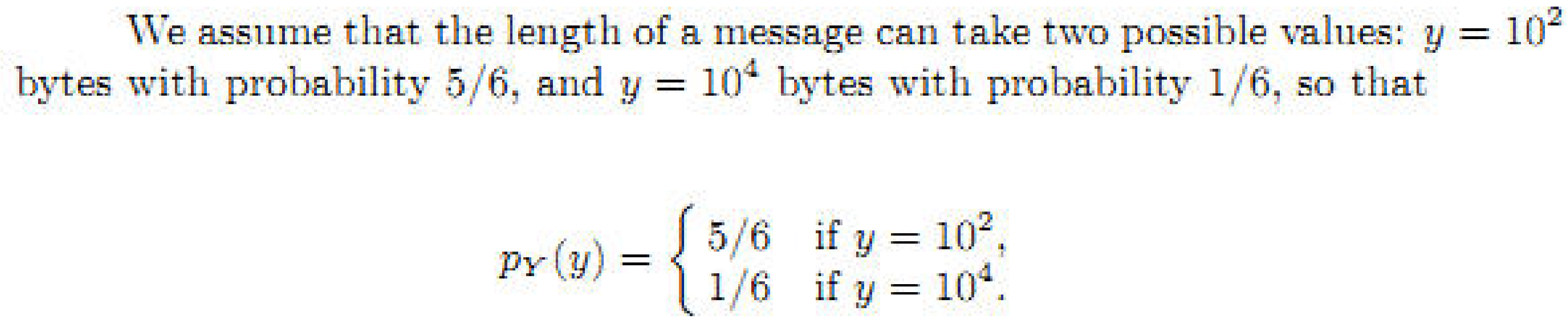


Example 2.13. Consider a transmitter that is sending messages over a computer network. Let us define the following two random variables:

X : the travel time of a given message, $\quad Y$: the length of the given message.

We know the PMF of the travel time of a message that has a given length, and we know the PMF of the message length. We want to find the (unconditional) PMF of the travel time of a message.

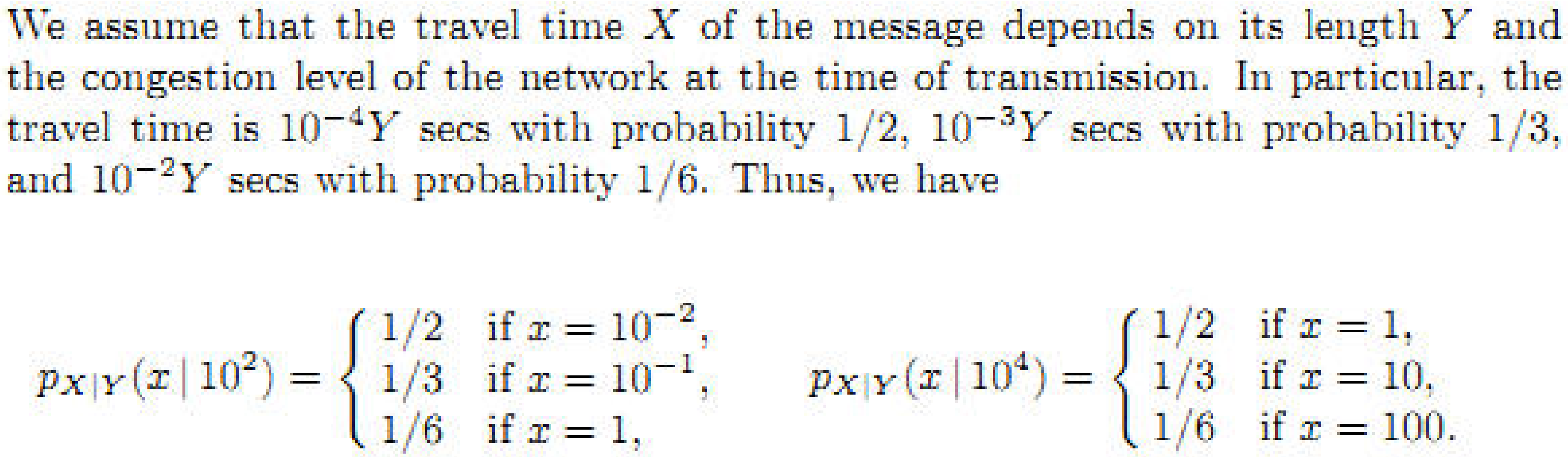
Example (cont.)



We assume that the length of a message can take two possible values: $y = 10^2$ bytes with probability $5/6$, and $y = 10^4$ bytes with probability $1/6$, so that

$$p_Y(y) = \begin{cases} 5/6 & \text{if } y = 10^2, \\ 1/6 & \text{if } y = 10^4. \end{cases}$$

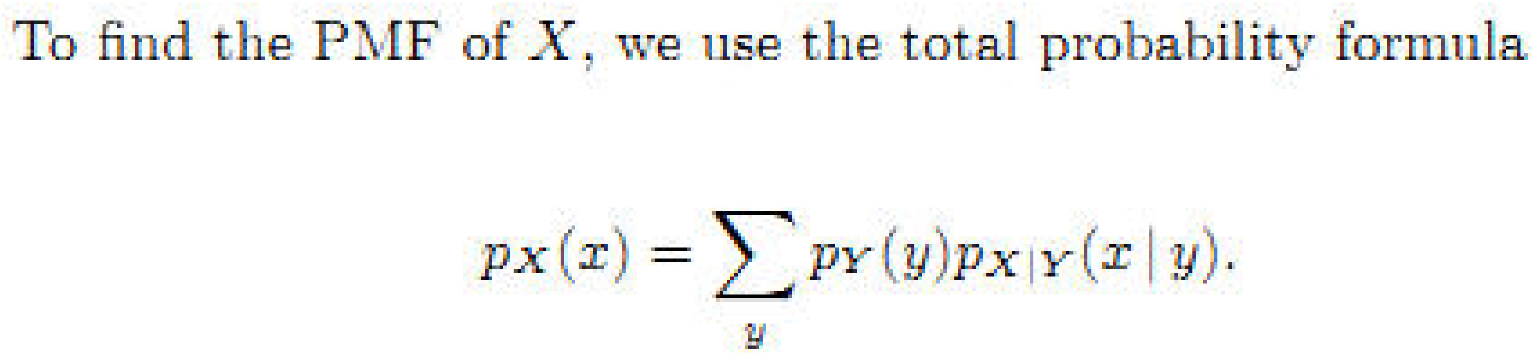
Example (cont.)



We assume that the travel time X of the message depends on its length Y and the congestion level of the network at the time of transmission. In particular, the travel time is $10^{-4}Y$ secs with probability $1/2$, $10^{-3}Y$ secs with probability $1/3$, and $10^{-2}Y$ secs with probability $1/6$. Thus, we have

$$p_{X|Y}(x\,|\,10^2) = \begin{cases} 1/2 & \text{if } x = 10^{-2}, \\ 1/3 & \text{if } x = 10^{-1}, \\ 1/6 & \text{if } x = 1, \end{cases} \qquad p_{X|Y}(x\,|\,10^4) = \begin{cases} 1/2 & \text{if } x = 1, \\ 1/3 & \text{if } x = 10, \\ 1/6 & \text{if } x = 100. \end{cases}$$

Example (cont.)



To find the PMF of X, we use the total probability formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \mid y).$$

Example (cont.)

We obtain

$$px(10^{-2}) = \frac{5}{6} \cdot \frac{1}{2}, \qquad px(10^{-1}) = \frac{5}{6} \cdot \frac{1}{3}, \qquad px(1) = \frac{5}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{2},$$

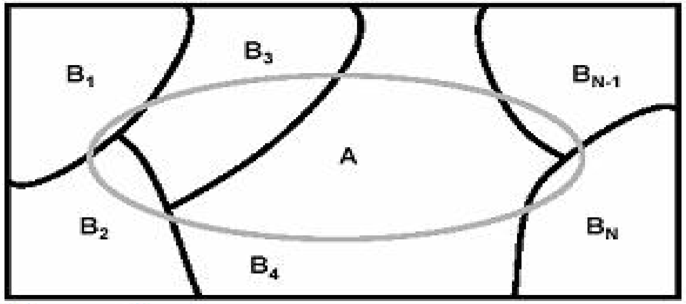
$$px(10) = \frac{1}{6} \cdot \frac{1}{3}, \qquad px(100) = \frac{1}{6} \cdot \frac{1}{6}.$$

Law of Total Probability

Let $B_{1, \ldots} B_N$ be N mutually exclusive events, whose union gives the sample space Ω. Hence the events B constitute a partition of Ω

Now consider an event A, a subset of Ω. This event can be represented as

$$A = A \cap \Omega = A \cap (B_1 \cup B_2 \cup \ldots \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_N)$$



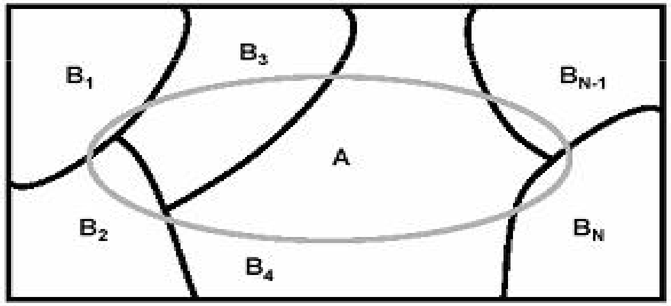
Since the B_i are mutually exclusive

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_N)$$

$$P(A) = P(A \mid B_1)P(B_1) + \ldots + P(A \mid B_N)P(B_N) = \sum_{k=1}^{N} P(A \mid B_k)P(B_k)$$

Bayes Rule

- We now pose the following question: Given that the event *A* has occurred. What is the probability that any single one of the event *B's* occur?



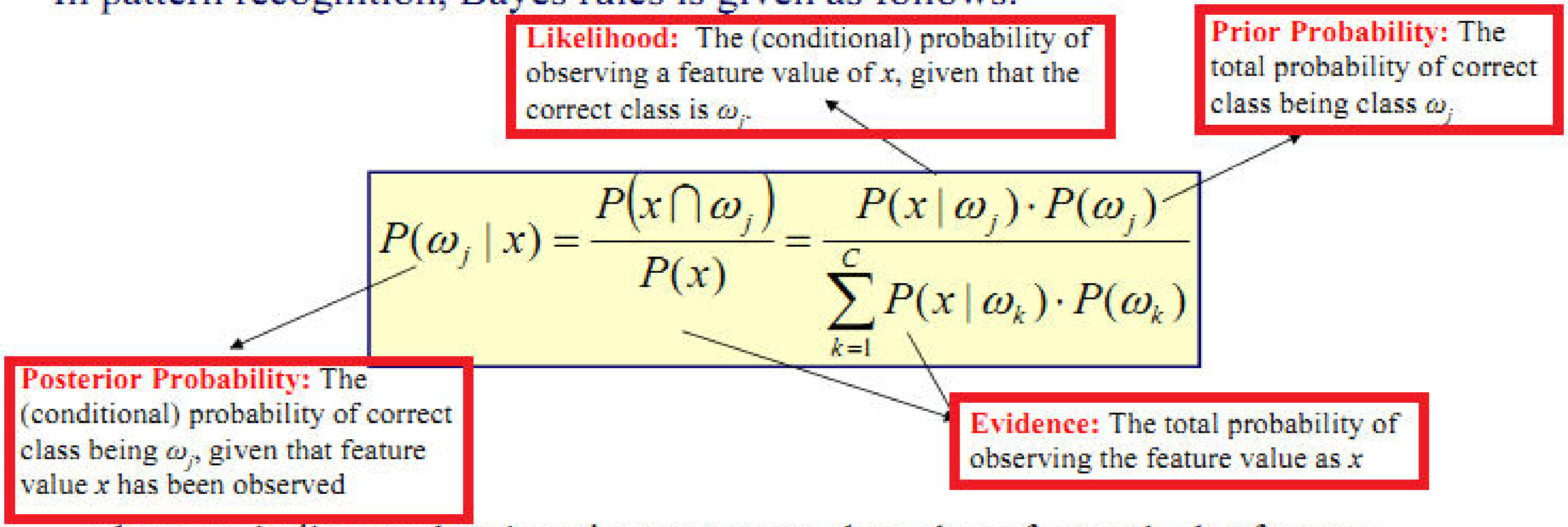
$$\boxed{P(B_j \mid A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{k=1}^{N} P(A \mid B_k) \cdot P(B_k)}}$$

This is known as the Bayes rule

Rev. Thomas Bayes,
(1702-1761)

Bayes Rule

In pattern recognition, Bayes rules is given as follows:



Likelihood: The (conditional) probability of observing a feature value of *x*, given that the correct class is ω_j.

Prior Probability: The total probability of correct class being class ω_j

$$P(\omega_j \mid x) = \frac{P(x \cap \omega_j)}{P(x)} = \frac{P(x \mid \omega_j) \cdot P(\omega_j)}{\sum_{k=1}^{C} P(x \mid \omega_k) \cdot P(\omega_k)}$$

Posterior Probability: The (conditional) probability of correct class being ω_j, given that feature value *x* has been observed

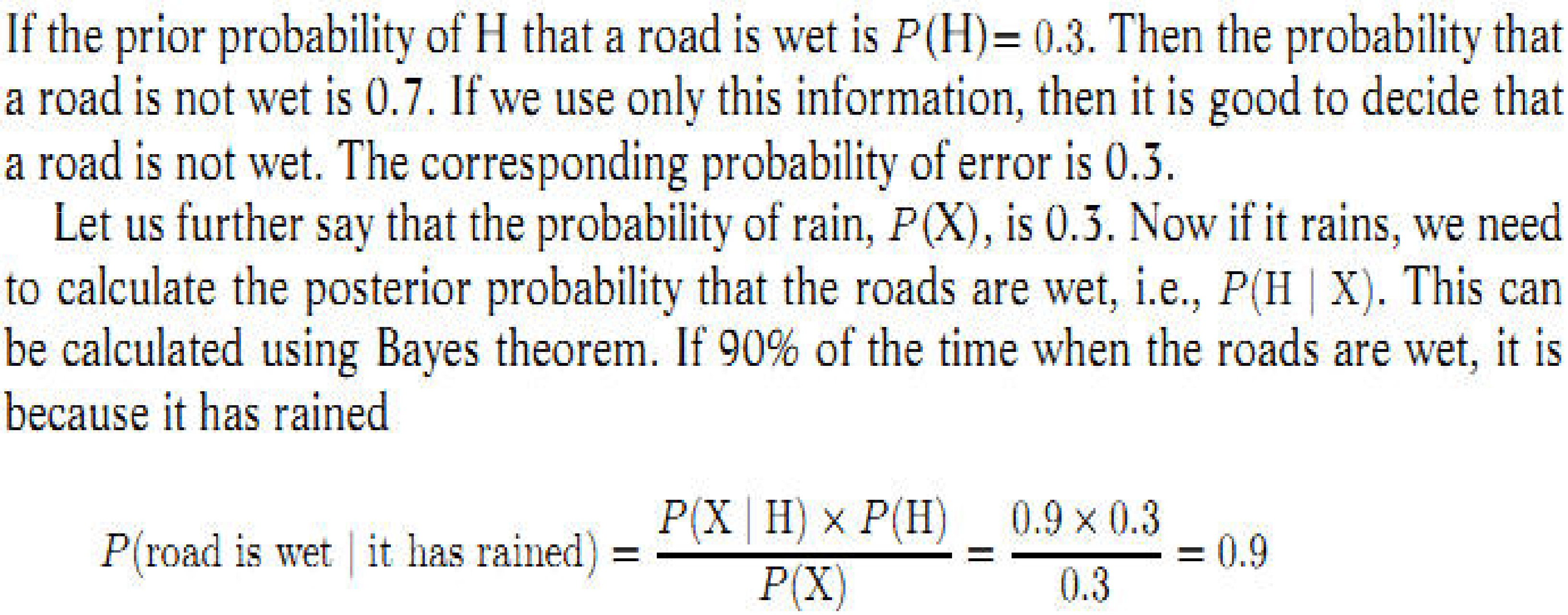
Evidence: The total probability of observing the feature value as *x*

where, ω_j indicates class j, and *x* represents the value of a particular feature.

A Bayes classifier, decides on the class ω_j that has the largest posterior probability.

The Bayes classifier is statistically the ***best*** classifier one can possibly construct.

Example



If the prior probability of H that a road is wet is $P(\mathrm{H}) = 0.3$. Then the probability that a road is not wet is 0.7. If we use only this information, then it is good to decide that a road is not wet. The corresponding probability of error is 0.3.

Let us further say that the probability of rain, $P(\mathrm{X})$, is 0.3. Now if it rains, we need to calculate the posterior probability that the roads are wet, i.e., $P(\mathrm{H} \mid \mathrm{X})$. This can be calculated using Bayes theorem. If 90% of the time when the roads are wet, it is because it has rained

$$P(\text{road is wet} \mid \text{it has rained}) = \frac{P(\mathrm{X} \mid \mathrm{H}) \times P(\mathrm{H})}{P(\mathrm{X})} = \frac{0.9 \times 0.3}{0.3} = 0.9$$

Many Dimensions

In most practical applications, we have more then one feature, and therefore the random variable x must be replaced with a *random vector* $\mathbf{x}$. $P(x)$ ➔ $P(\mathbf{x})$

The joint probability mass function $P(\mathbf{x})$ still satisfies the axioms of probability

The Bayes rule is then

$$\boxed{P(\omega_j \mid \mathbf{x}) = \frac{P\left(\mathbf{x} \cap \omega_j\right)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid \omega_j) \cdot P(\omega_j)}{\sum_{k=1}^{C} P(\mathbf{x} \mid \omega_k) \cdot P(\omega_k)}}$$

While the notation changes only slightly, the implications are quite substantial\

- The curse of dimensionality

Random Vectors

- A random vector is always represented as a column vector: $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}}$
- The joint cumulative and probability mass/density functions are defined as:

$$F_{\mathbf{x}}(\mathbf{x}) = P[(X_1 \le x_1) \cap (X_2 \le x_2) \cap \ldots \cap (X_d \le x_d)]$$
$$P_{\mathbf{x}}(\mathbf{x}) = \frac{\Delta^d F_{\mathbf{x}}(\mathbf{x})}{\Delta x_1 \Delta x_2 \ldots \Delta x_d}$$

$$f_{\mathbf{x}}(\mathbf{x}) = P[(X_1 \le x_1) \cap (X_2 \le x_2) \cap \ldots \cap (X_d \le x_d)]$$
$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\partial^d f_{\mathbf{x}}(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_d}$$

- To obtain the cdf/pdf of a subset of the variables, *marginal pdf/cdf*, the variables that are of not interest are *integrated out*.
- Similar expectation and covariance operations can be computed on random vectors

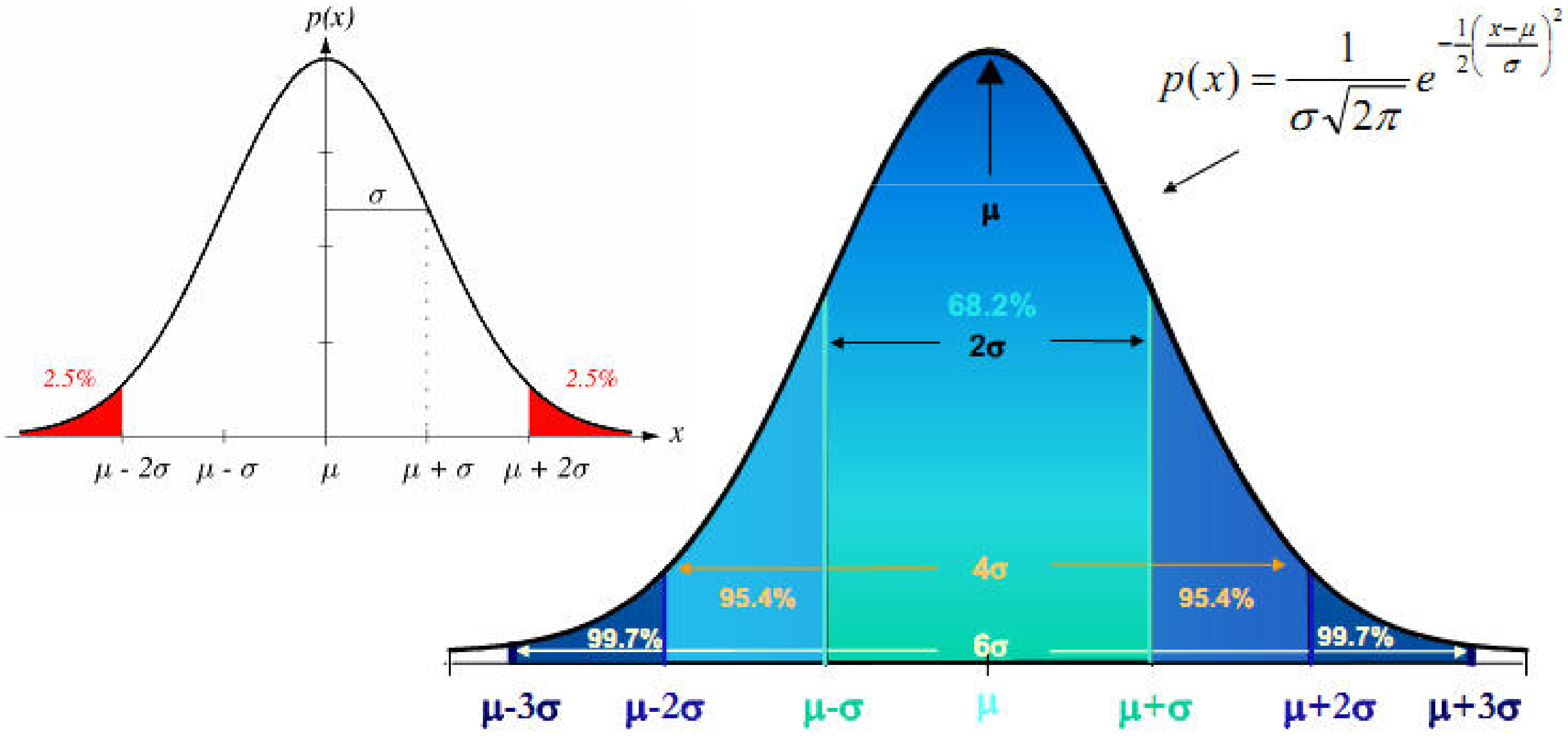
$$\varepsilon[\mathbf{x}] = \begin{bmatrix} \varepsilon(x_1) \\ \varepsilon(x_2) \\ \varepsilon(x_3) \\ \vdots \\ \varepsilon(x_d) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_d \end{bmatrix}$$

$$\mathbf{\Sigma} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] = \begin{bmatrix} E\left[(x_1-\mu_1)(x_1-\mu_1)^t\right] & E\left[(x_1-\mu_1)(x_d-\mu_d)^t\right] \\ E\left[(x_d-\mu_d)(x_1-\mu_1)^t\right] & E\left[(x_d-\mu_d)(x_d-\mu_d)^t\right] \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$$

If the x_i's are statistically independent, the off-diagonal elements of the $\sum$ will be zero.

Gaussian Distribution

- By far the most important and most commonly observed (cont.) probability distribution



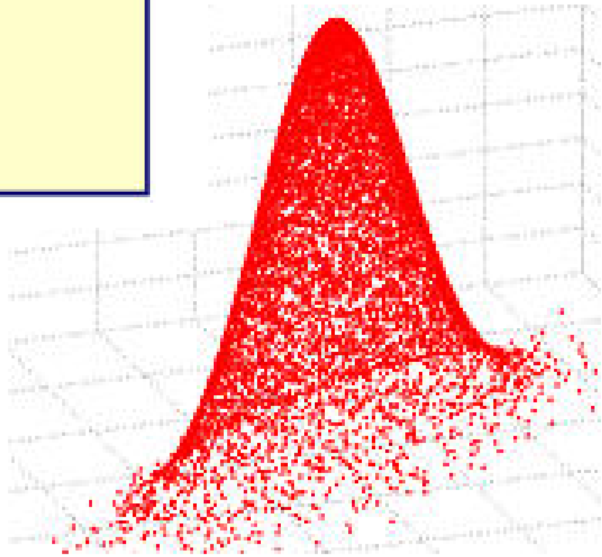
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Multivariate Gaussian Distribution

- In *d-dimensional* space, the Gaussian pdf is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}$$

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$$



Linear Algebra

- **x, u, v, w (bold face, lower case)** d-dimensional column vector
- **xi** ith element of the vector
- **X (bold face upper case) dxk** dimensional matrix

Linear Algebra

- **A linear combination of vectors is another vector:**

$$v = a_1 u_1 + a_2 u_2 + \cdots + a_n u_n$$

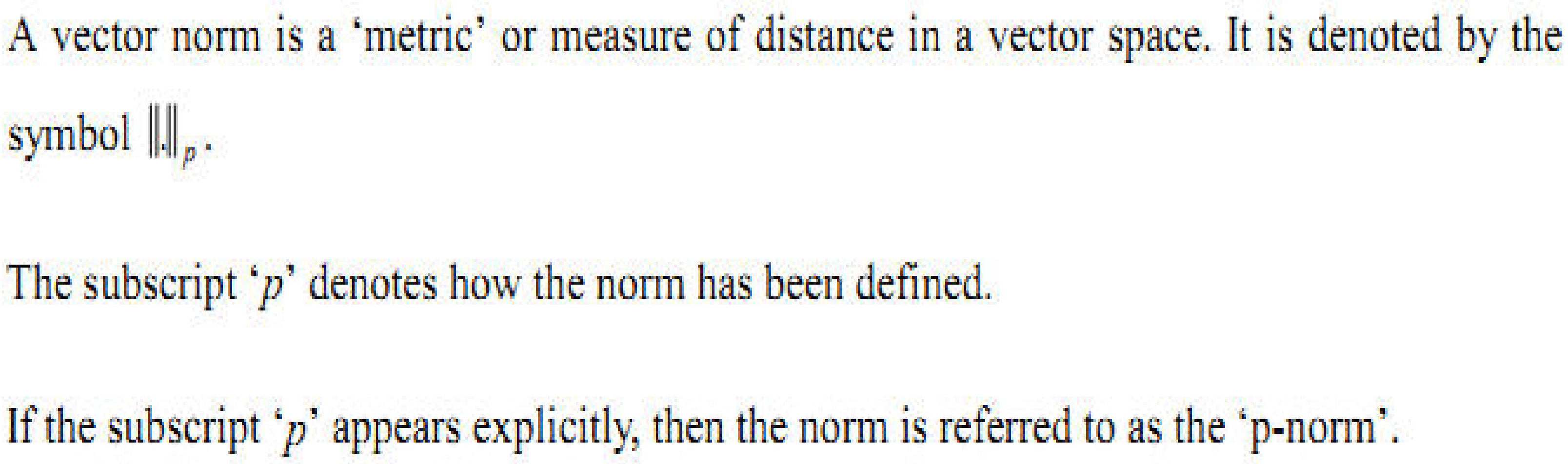
- A collection of vectors are **linearly dependent**, if any one of them can be written as a linear combination of others with at least one non-zero scalar.

Linear Algebra

- A basis of V is a collection of linearly independent vectors such that any vector **v** in V can be written as a linear combination of these basis vectors.
- That if **B={u1,u2,…,un}** is a basis for V, than any **v** in V can be written as

$$v = a_1 u_1 + a_2 u_2 + \cdots + a_n u_n$$

Linear Algebra



A vector norm is a ‘metric’ or measure of distance in a vector space. It is denoted by the symbol $\|.\|_p$.

The subscript ‘p’ denotes how the norm has been defined.

If the subscript ‘p’ appears explicitly, then the norm is referred to as the ‘p-norm’.

Linear Algebra

For any n-dimensional vector $\mathbf{x} = \left[x_1, x_2, x_3, \cdots, x_n\right]^T$, any norm of $\mathbf{x}$ must satisfy the following properties:

1) $\|\mathbf{x}\|_p \geq 0$

2) $\|\mathbf{x}\|_p = 0$ *iff* $\mathbf{x} = \mathbf{0}$, i.e. $\mathbf{x}$ is the 'null' vector with $x_i = 0, \quad i = 1,2,3,\ldots,n$

3) For any scalar α, $\|\alpha\mathbf{x}\|_p = |\alpha|.\|\mathbf{x}\|_p$

4) If $\mathbf{y}$ is another n-dimensional vector, then $\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. This relationship is known as the 'triangle inequality'.

Linear Algebra

$$\|\mathbf{x}\|_2 = \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2}$$

$$\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$$

Examples:

Let $\mathbf{x} = [-1, 1, -2]^T$. Then,

$$\|\mathbf{x}\|_2 = \sqrt{(-1)^2 + 1^2 + (-1)^2} = \sqrt{6}$$

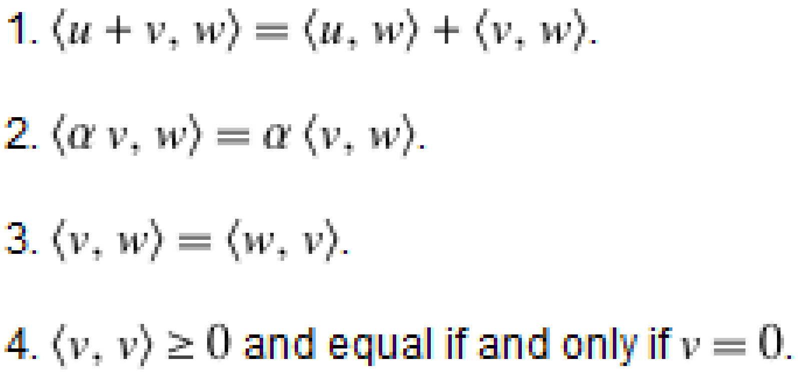
while

$$\|\mathbf{x}\|_\infty = \max\{|-1|, |1|, |-2|\} = 2$$

Linear Algebra

•An inner product in a vector space, is a way to multiply vectors together, with the result of this multiplication being a scalar.

•More precisely, for a real vector space, an inner product satisfies the following four properties:



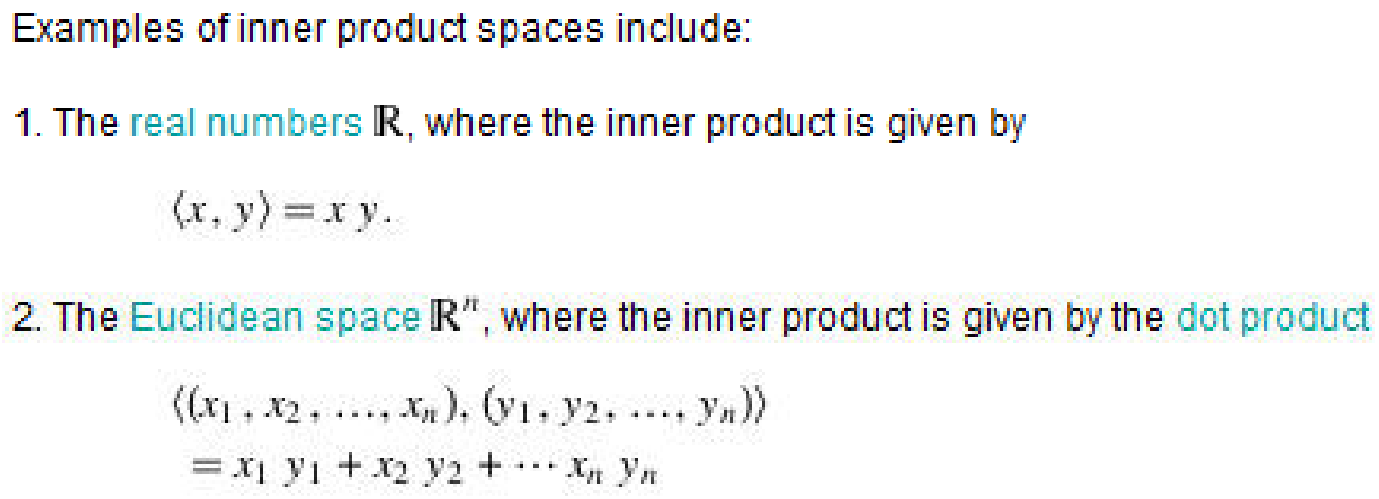
1. $\langle u+v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$.

2. $\langle \alpha\, v, w\rangle = \alpha\, \langle v, w\rangle$.

3. $\langle v, w\rangle = \langle w, v\rangle$.

4. $\langle v, v\rangle \geq 0$ and equal if and only if $v = 0$.

Linear Algebra



Examples of inner product spaces include:

1. The real numbers $\mathbb{R}$, where the inner product is given by

$$\langle x, y \rangle = x\, y.$$

2. The Euclidean space $\mathbb{R}^n$, where the inner product is given by the dot product

$$\begin{aligned}&\langle (x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \rangle \\ &= x_1\, y_1 + x_2\, y_2 + \cdots x_n\, y_n\end{aligned}$$

Linear Algebra

Orthogonality: Two vectors, *x* and *y*, in an inner product space, *V*, are *orthogonal* if their inner product, is zero.

Linear Algebra

Gradient:

$$\nabla f(\mathbf{x}) = grad\ f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \partial f(\mathbf{x})/\partial x_1 \\ \partial f(\mathbf{x})/\partial x_2 \\ \vdots \\ \partial f(\mathbf{x})/\partial x_d \end{pmatrix}$$

Jacobian:

$$J(\mathbf{f}(\mathbf{x})) = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right) = \begin{pmatrix} \partial f_1(\mathbf{x})/\partial x_1 & \cdots & \partial f_1(\mathbf{x})/\partial x_d \\ \vdots & \ddots & \vdots \\ \partial f_n(\mathbf{x})/\partial x_1 & \cdots & \partial f_n(\mathbf{x})/\partial x_d \end{pmatrix}$$

Linear Algebra

Hessian:

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Linear Algebra

Taylor expansions for vector valued functions:

$$y = f(\mathbf{x} + \Delta\mathbf{x}) \approx f(\mathbf{x}) + J(\mathbf{x})\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^{\mathrm{T}} H(\mathbf{x})\Delta\mathbf{x}$$