Review of Probability Axioms and Laws

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

1. D. P. Bertsekas, J. N. Tsitsiklis, “Introduction to Probability,” Athena Scientific, 2008.

What is "Probability" ?

- Probability was developed to describe phenomena that cannot be predicted with certainty
  - Frequency of occurrences
  - Subjective beliefs

- Everyone accepts that the probability (of a certain thing to happen) is a number between 0 and 1 (?)

- Measures deduced from probability axioms and theories (laws/rules) can help us deal with and quantify "information"

Sets (1/2)

- A **set** is a collection of objects which are the **elements** of the set
  - If x is an element of set S, denoted by $x \in S$
  - Otherwise denoted by $x \notin S$
- A set that has no elements is called **empty set** is denoted by Ø
- Set specification
  - Countably finite: $\{1,2,3,4,5,6\}$
  - Countably infinite: $\{0,2,-2,4,-4,\ldots\}$
  - With a certain property: $\{k \mid k/2 \text{ is integer}\}$

$$\{x \mid 0 \le x \le 1\}$$

$$\{x \mid x \text{ satisfies } P\}$$

such that

Sets (2/2)

- If every element of a set S is also an element of a set T, then S is a **subset** of T
  - Denoted by $S \subset T$ or $T \supset S$
- If $S \subset T$ and $T \subset S$, then the two sets are **equal**
  - Denoted by $S = T$
- The universal set, denoted by Ω, which contains all objects of interest in a particular context
  - After specifying the context in terms of universal set Ω, we only consider sets S that are subsets of Ω

Set Operations (1/3)

- Complement
  - The **complement** of a set S with respect to the universe Ω, is the set $\{x \in \Omega \mid x \notin S\}$, namely, the set of all elements that do not belong to S, denoted by S^c
  - The complement of the universe $\Omega^c = \varnothing$

- Union
  - The **union** of two sets S and T is the set of all elements that belong to S or T, denoted by $S \bigcup T$

    $$S \bigcup T = \{x \mid x \in S \text{ or } x \in T\}$$

- Intersection
  - The **intersection** of two sets S and T is the set of all elements that belong to both S and T, denoted by $S \bigcap T$

    $$S \bigcap T = \{x \mid x \in S \text{ and } x \in T\}$$

Set Operations (2/3)

- The union or the intersection of several (or even infinite many) sets

$$\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \cdots = \{x \mid x \in S_n \text{ for some } n\}$$

$$\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \cdots = \{x \mid x \in S_n \text{ for all } n\}$$

- Disjoint
  - Two sets are **disjoint** if their intersection is empty (e.g., $S \cap T$ = Ø)

- 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

Set Operations (3/3)

- Visualization of set operations with Venn diagrams

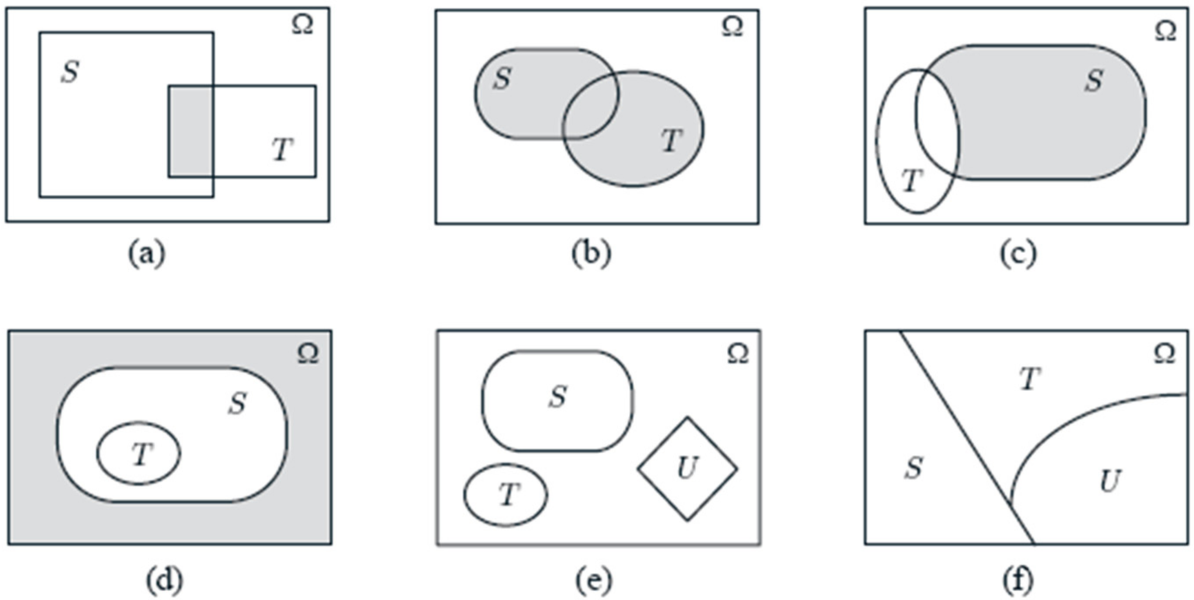


Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of the set Ω.

The Algebra of Sets

- The following equations are the elementary consequences of the set definitions and operations

commutative

$$S \cup T = T \cup S,$$

associative

$$S \cup (T \cup U) = (S \cup T) \cup U$$

distributive

$$S \cap (T \cup U) = (S \cap T) \cup (S \cap U),$$

distributive

$$S \cup (T \cap U) = (S \cup T) \cap (S \cup U),$$

$$(S^c)^c = S,$$

$$S \cap S^c = \emptyset$$

$$S \cup \Omega = \Omega,$$

$$S \cap \Omega = S.$$

- De Morgan's law

$$\left(\bigcup_n S_n\right)^c = \bigcap_n S_n^c \qquad \left(\bigcap_n S_n\right)^c = \bigcup_n S_n^c$$

Probabilistic Models (1/2)

- A probabilistic model is a mathematical description of an uncertainty situation
  - It has to be in accordance with a fundamental framework to be discussed shortly

- 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 A

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

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

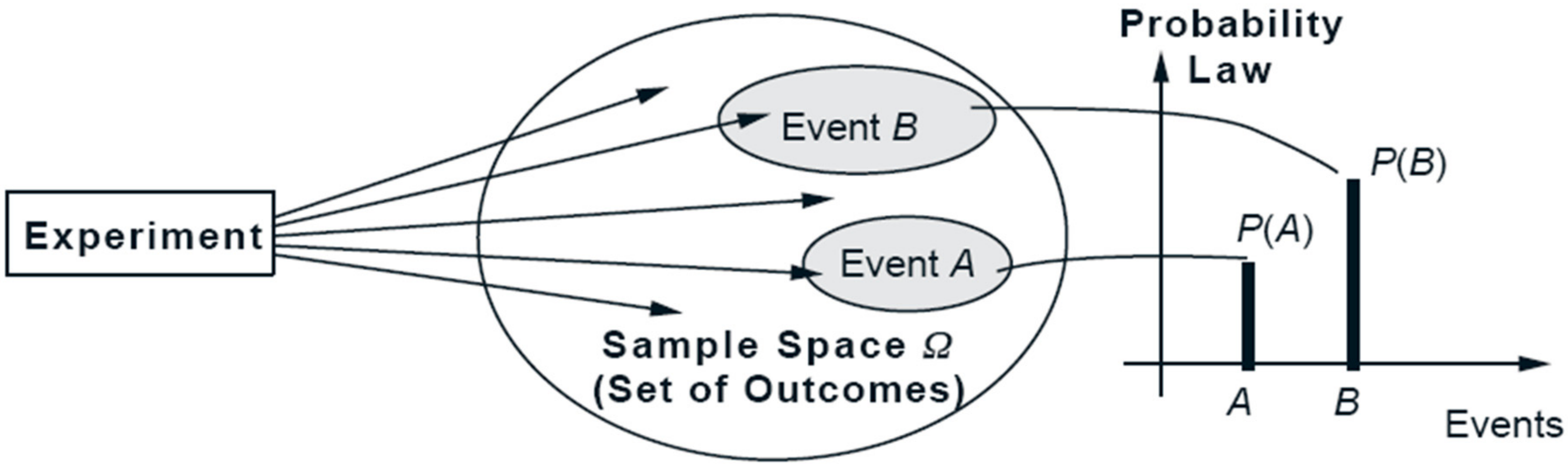
   Furthermore, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

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

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

Probabilistic Models (2/2)

- The main ingredients of a probabilistic model



Sample Spaces and Events

- Each probabilistic model involves an underlying process, called the **experiment**
  - That 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.

Sample Spaces and Events (2/2)

- Properties of the sample space
  - Elements of the sample space must be **mutually exclusive**
  - The sample space must be **collectively exhaustive**
  - The sample space should be at the "right" granularity (avoiding irrelevant details)

Probability Laws

- Discrete Probability Law
  - If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \ldots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}(\{s_1, s_2, \ldots, s_n\}) = \mathbf{P}(\{s_1\}) + \mathbf{P}(\{s_2\}) + \cdots + \mathbf{P}(\{s_n\})$$
$$= \mathbf{P}(s_1) + \mathbf{P}(s_2) + \cdots + \mathbf{P}(s_n)$$

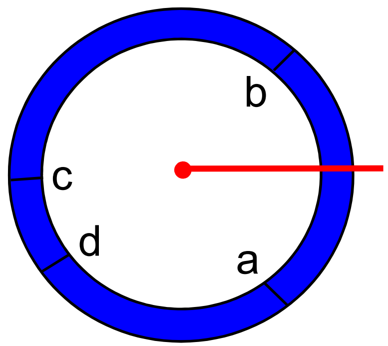
- Discrete Uniform Probability Law
  - If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{number of element of } A}{n}$$

Continuous Models

- Probabilistic models with continuous sample spaces
  - It is inappropriate to assign probability to each single-element event (?)
  - Instead, it makes sense to assign probability to any interval (one-dimensional) or area (two-dimensional) of the sample space

- Example: Wheel of Fortune



$$\mathbf{P}(\{0.3\}) = ?$$

$$\mathbf{P}(\{0.33\}) = ?$$

$$\mathbf{P}(\{0.333\}) = ?$$

$$\cdots$$

$$\mathbf{P}(\{x | a \le x \le b\}) = ?$$

Properties of Probability Laws

- Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below

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

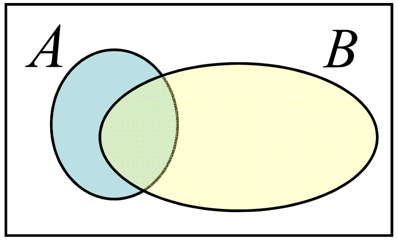
(c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

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

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?

Conditional Probabilities Satisfy the Three Axioms

- Nonnegative:

$$\mathbf{P}(A|B) \geq 0$$

- Normalization:

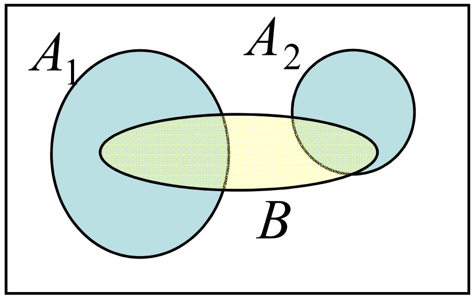
$$\mathbf{P}(\Omega|B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1$$

- Additivity:If A_1 and A_2 are two disjoint events

$$\begin{aligned} \mathbf{P}(A_1 \cup A_2 | B) &= \frac{\mathbf{P}((A_1 \cup A_2) \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}((A_1 \cap B) \cup (A_2 \cap B))}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1 \cap B) + \mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)} \\ &= \mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) \end{aligned}$$

distributive

disjoint sets



Multiplication (Chain) Rule

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

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbf{P}(A_1)\mathbf{P}(A_2 | A_1)\mathbf{P}(A_3 | A_1 \cap A_2)\cdots\mathbf{P}\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right)$$

  – The above formula can be verified by writing

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

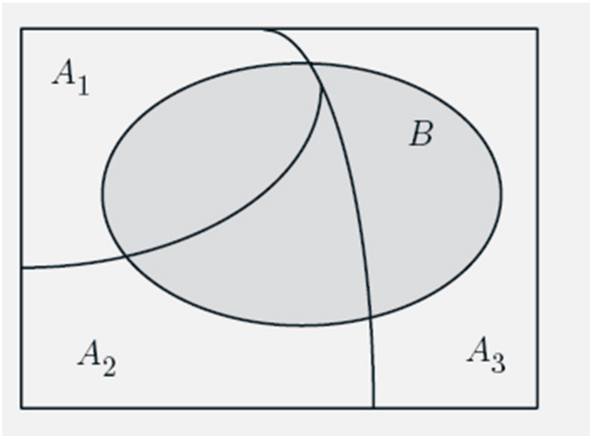
  – For the case of just two events, the multiplication rule is simply the definition of conditional probability

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

Total Probability Theorem

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

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



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

Bayes' Rule

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

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

Multiplication rule

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

Total probability theorem

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

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

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)

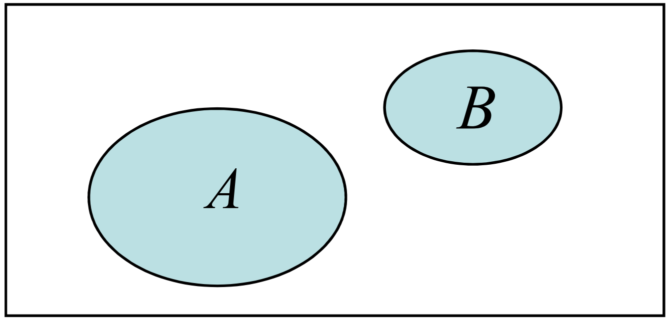
$$\Rightarrow \mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \mathbf{P}(A)$$

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

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

Conditional Independence (1/2)

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

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

  – We also know that

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

multiplication rule

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

  – 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

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \quad \not\Leftrightarrow \quad \mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$$

- If A and B are independent, the same holds for
  (i) A and B^c
  (ii) A^c and B
  (iii) A^c and B^c

Independence of a Collection of Events

- We say that the events $A_1, A_2, \ldots, A_n$ are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i), \text{ for every subset } S \text{ of } \{1,2,\ldots,n\}$$

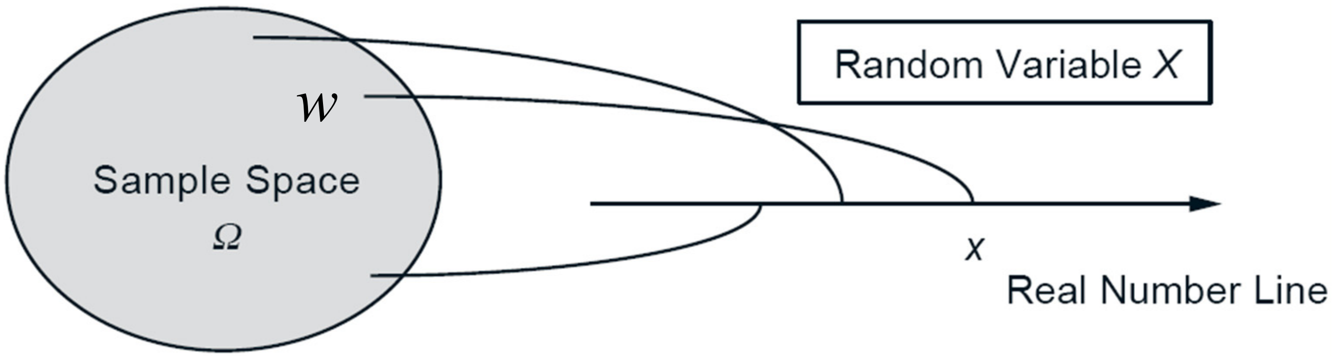
- For example, the independence of three events A_1, A_2, A_3 amounts to satisfying the four conditions

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2)$$
$$\mathbf{P}(A_1 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_3)$$
$$\mathbf{P}(A_2 \cap A_3) = \mathbf{P}(A_2)\mathbf{P}(A_3)$$
$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3)$$

$2^n - n - 1$

Random Variables

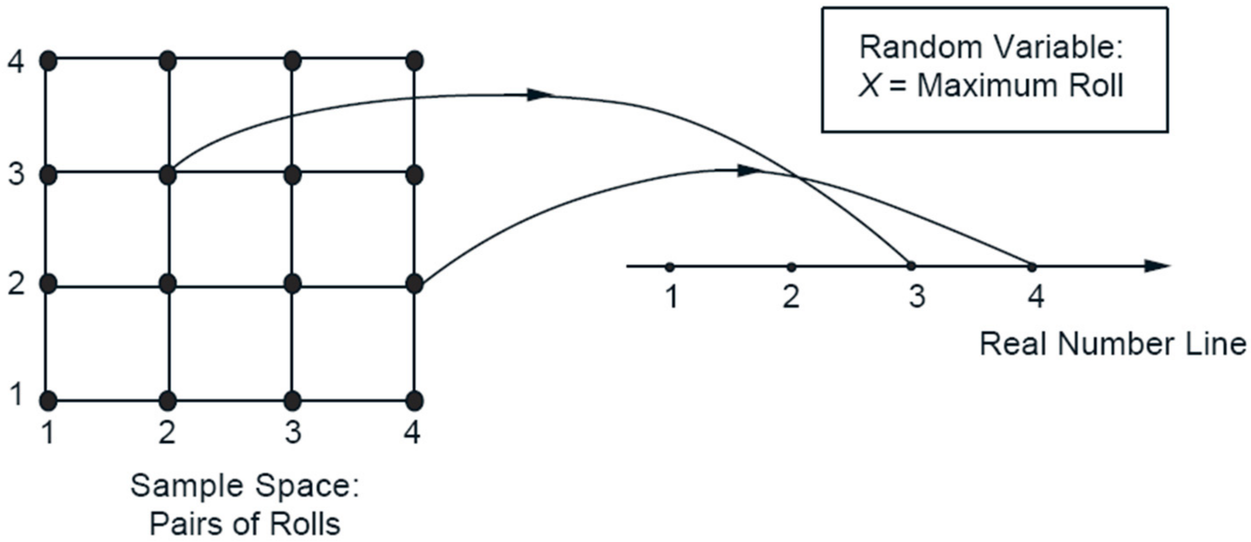
- Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome
  - This number is referred to as the (numerical) value of the random variable
  - We can say a random variable is a real-valued function of the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



$$X : w \rightarrow x$$

Random Variables: Example

- An experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls
  - If the outcome of the experiment is (4, 2), the value of this random variable is 4
  - If the outcome of the experiment is (3, 3), the value of this random variable is 3



  - Can be one-to-one or many-to-one mapping

Discrete/Continuous Random Variables

- 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

Concepts Related to Discrete Random Variables

- For a probabilistic model of an experiment
  - A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values
  - A (discrete) random variable has an associated **probability mass function** (PMF), which gives the probability of each numerical value that the random variable can take
  - A **function of a random variable** defines another random variable, whose PMF can be obtained from the PMF of the original random variable

Probability Mass Function

- A (discrete) random variable X is characterized through the probabilities of the values that it can take, which is captured by the probability mass function (PMF) of X, denoted $p_X(x)$

$$p_X(x) = \mathbf{P}(\{X = x\}) \text{ or } p_X(x) = \mathbf{P}(X = x)$$

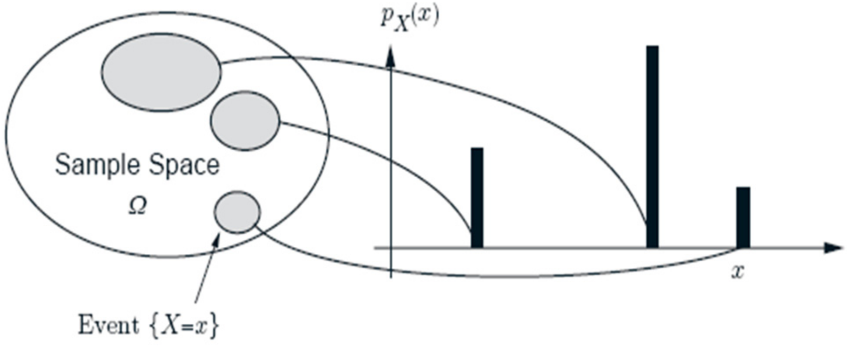
  - The sum of probabilities of all outcomes that give rise to a value of X equal to x
  - **Upper case** characters (e.g., X) denote random variables, while **lower case** ones (e.g., x) denote the numerical values of a random variable
- The summation of the outputs of the PMF function of a random variable over all it possible numerical values is equal to one $\sum_x p_X(x) = 1$

$\{X = x\}$'s are disjoint and form a partition of the sample space

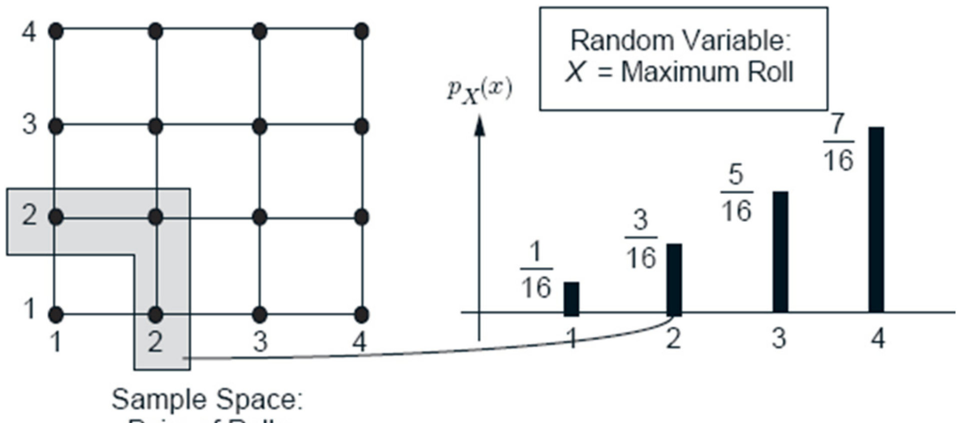
Calculation of the PMF

- For each possible value x of a random variable X :
  1. Collect all the possible outcomes that give rise to the event $\{X = x\}$
  2. Add their probabilities to obtain $p_X(x)$

- An example: the PMF $p_X(x)$ of the random variable X = maximum roll in two independent rolls of a fair 4-sided die



(a)



(b)

Expectation

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

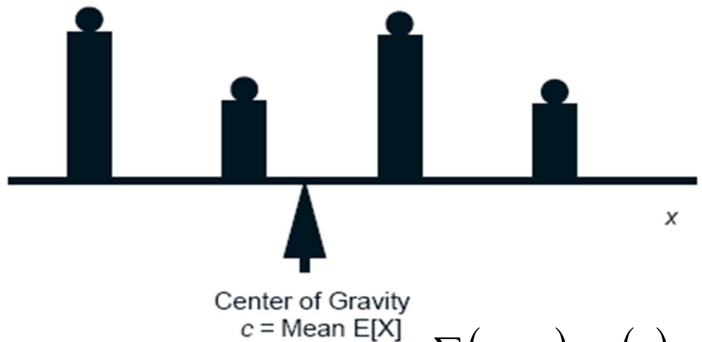
$$\mathbf{E}[X] = \sum_x x p_X(x)$$

  – 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

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

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

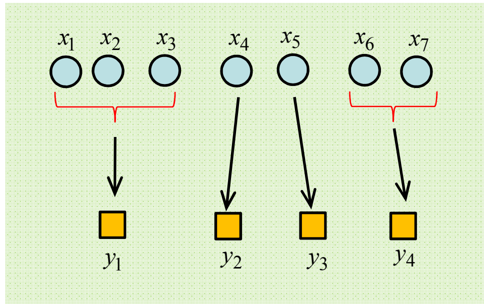


$$\sum_x (x-c) p_X(x) = 0$$

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

Expectations for Functions of 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 $p_Y(y) = \sum_{\{x \mid g(x) = y\}} p_X(x)$

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

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

$$= \sum_x g(x) p_X(x)$$

?

Variance

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

$$\operatorname{var}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$$
$$= \sum_x (x - \mathbf{E}[X])^2 \, p_X(x)$$

  - 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)

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

    - Easier to interpret, because it has the same units as X

Properties of Mean and Variance

- Let X be a random variable and let

$$Y = aX + b$$ a linear function of X

where a and b are given scalars

Then,

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

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

- If $g(X)$ is a linear function of X, then

$$\mathbf{E}[g(X)] = g(\mathbf{E}[X])$$ How to verify it ?

Joint PMF of Random Variables

- Let X and Y be random variables associated with the same experiment (also the same sample space and probability laws), the **joint PMF** of X and Y is defined by

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

- if event A is the set of all pairs (x,y) that have a certain property, then the probability of A can be calculated by

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

  – Namely, A can be specified in terms of X and Y

Marginal PMFs of Random Variables

- The **PMFs** of random variables X and Y can be calculated from their **joint PMF**

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

  - $p_X(x)$ and $p_Y(y)$ are often referred to as the **marginal PMFs**
  - The above two equations can be verified by

$$\begin{aligned} p_X(x) &= \mathbf{P}(X=x) \\ &= \sum_y \mathbf{P}(X=x, Y=y) \\ &= \sum_y p_{X,Y}(x,y) \end{aligned}$$

Conditioning

- Recall that conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information

- In the same spirit, we can define **conditional PMFs**, given the occurrence of a certain event or given the value of another random variable

Conditioning a Random Variable on an Event (1/2)

- The **conditional PMF** of a random variable X , conditioned on a particular event A with $\mathbf{P}(A) > 0$, is defined by (where X and A are associated with the same experiment)

$$P_{X|A}(x) = \mathbf{P}(X = x|A) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}$$

- Normalization Property
  - Note that the events $\mathbf{P}(\{X = x\} \cap A)$ are disjoint for different values of X, their union is A

$$\mathbf{P}(A) = \sum_x \mathbf{P}(\{X = x\} \cap A)$$

Total probability theorem

$$\therefore \sum_x P_{X|A}(x) = \sum_x \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)} = \frac{\sum_x \mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A)}{\mathbf{P}(A)} = 1$$

Conditioning a Random Variable on an Event (2/2)

- A graphical illustration

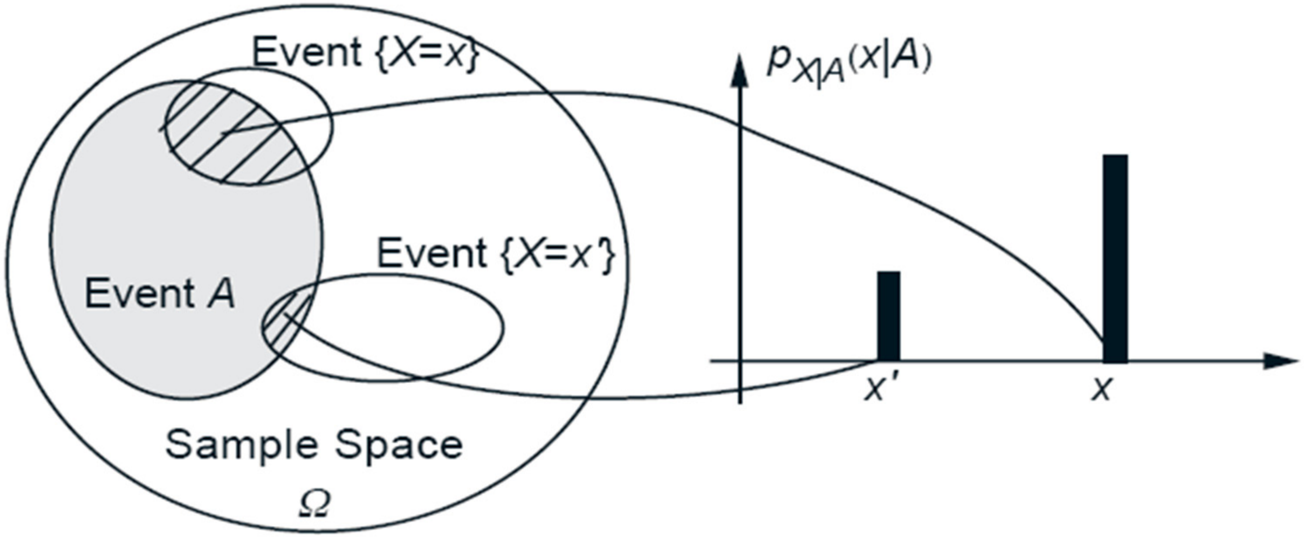


Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x, we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$ and normalize by diving with $\mathbf{P}(A)$.

$P_{X|A}(x)$ Is obtained by adding the probabilities of the outcomes that give rise to $X = x$ and be long to the conditioning event A

Conditioning a Random Variable on Another (1/2)

- Let X and Y be two random variables associated with the same experiment. The conditional PMF $p_{X|Y}$ of X given Y is defined as

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

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

Y is fixed on some value y

- Normalization Property $\sum_x p_{X|Y}(x|y) = 1$

- The conditional PMF is often convenient for the calculation of the joint PMF

multiplication (chain) rule

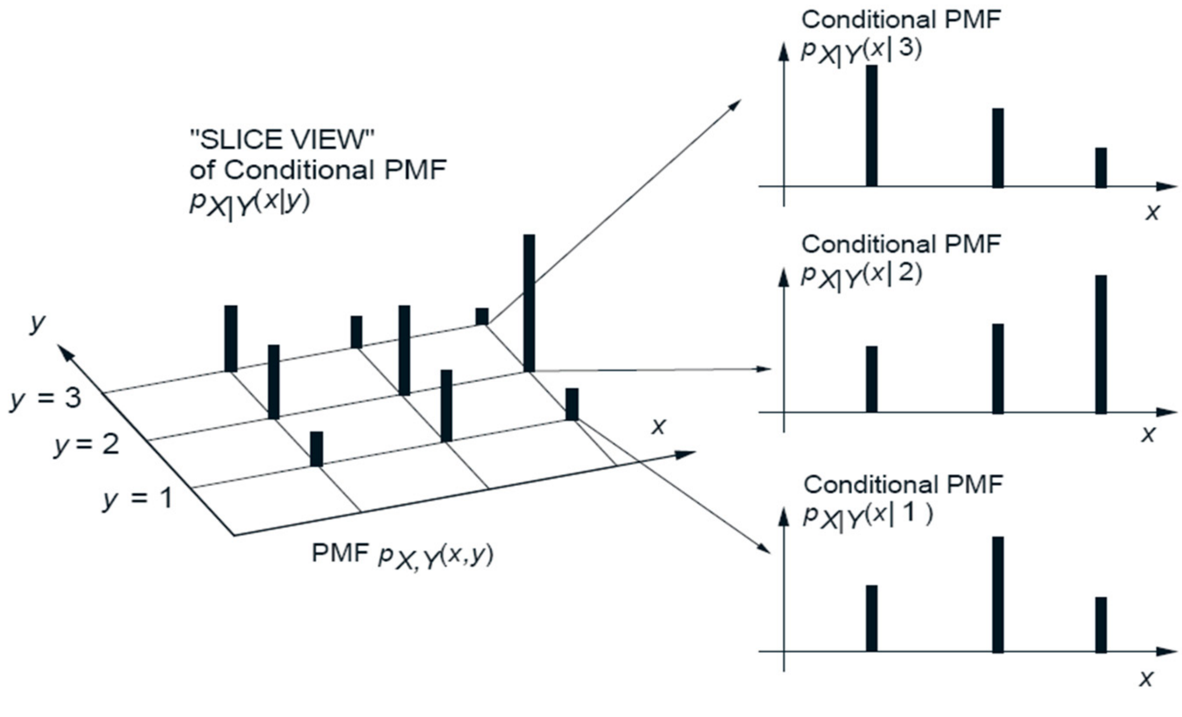
$$p_{X,Y}(x,y) = p_Y(y) p_{X|Y}(x|y) \ \left(= p_X(x) p_{Y|X}(y|x)\right)$$

Conditioning a Random Variable on Another (2/2)

- The conditional PMF can also be used to calculate the marginal PMFs

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

- Visualization of the conditional PMF $p_{X|Y}$



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

Independence of a Random Variable from an Event

- A random variable X is **independent of an event** A if

$$\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x)\mathbf{P}(A), \quad \text{for all } x$$

  - Require two events $\{X = x\}$ and A be independent for all x
- If a random variable X is **independent of an event** A and $\mathbf{P}(A) > 0$

$$\begin{aligned} p_{X|A}(x) &= \frac{\mathbf{P}(X = x \text{ and } A)}{\mathbf{P}(A)} \\ &= \frac{\mathbf{P}(X = x)\mathbf{P}(A)}{\mathbf{P}(A)} \\ &= \mathbf{P}(X = x) \\ &= p_X(x), \quad \text{for all } x \end{aligned}$$

Independence of Random Variables (1/2)

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

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

$$\text{or} \quad \mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x)\mathbf{P}(Y = y), \quad \text{for all } x, y$$

- If a random variable X is **independent of an random variable** Y

$$p_{X|Y}(x|y) = p_X(x), \quad \text{for all } y \text{ with } p_Y(y) > 0 \text{ all } x$$

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{X,Y}(x, y)}{p_Y(y)} \\ &= \frac{p_X(x) p_Y(y)}{p_Y(y)} \\ &= p_X(x), \quad \text{for all } y \text{ with } p(y) > 0 \text{ and all } x \end{aligned}$$

Independence of Random Variables (2/2)

- Random variables X and Y are said to be **conditionally independent**, given a positive probability event A, if

$$p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y), \quad \text{for all } x, y$$

  – Or equivalently,

$$p_{X|Y,A}(x|y) = p_{X|A}(x), \quad \text{for all } y \text{ with } p_{Y|A}(y) > 0 \text{ and all } x$$

- Note here that, as in the case of events, conditional independence may not imply unconditional independence and vice versa

Entropy (1/3)

- Three interpretations for quantity of information
  1. The amount of **uncertainty** before seeing an event
  2. The amount of **surprise** when seeing an event
  3. The amount of **information** after seeing an event

- The definition of information:

$$\textit{define} \quad 0\log_2 0 = 0$$

$$I(x_i) = \log_2 \frac{1}{P(x_i)} = -\log_2 P(x_i)$$

  - $P(x_i)$ the probability of an event x_i

- Entropy: the average amount of information

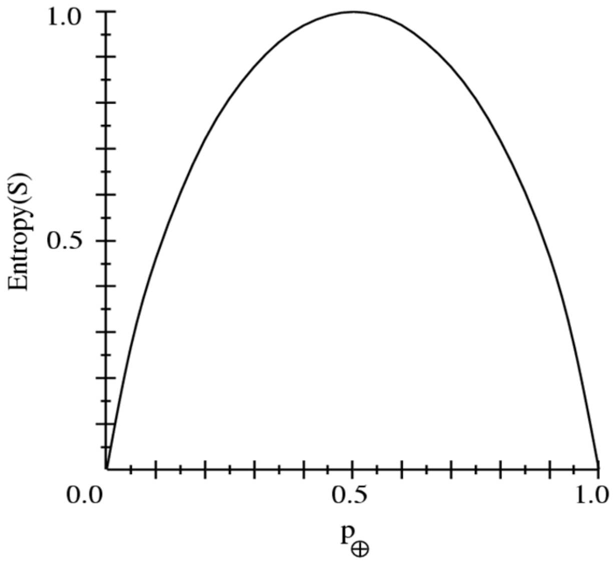
$$H(X) = E[I(X)]_X = E[-\log_2 P(x_i)]_X = \sum_{x_i} -P(x_i) \cdot \log_2 P(x_i)$$

where $X = \{x_1, x_2, \ldots, x_i, \ldots\}$

  - Have maximum value when the probability (mass) function is a uniform distribution

Entropy (2/3)

- For Boolean classification (0 or 1)



$$P_X(x) = \begin{cases} p_1, & x = 1 \\ p_2 = 1 - p_1, & x = 0 \end{cases}$$

$$Entropy\,(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

- Entropy can be expressed as the minimum number of bits of information needed to encode the classification of an arbitrary number of examples
  – If c classes are generated, the maximum of entropy can be

$$Entropy\,(X) = \log_2 c$$