

Foundations and Trends® in Computation  
Vol. XX, No. XX (2024) 1–221  
© 2024 L. Mahto  
DOI: 10.1561/XXXXXXXXXX

## **Introduction to probability and statistics: a computational framework of randomness**

Lakshman Mahto  
Assistant Professor (Mathematics), IIIT Dharwad, India  
[lakshman@iiitdwd.ac.in](mailto:lakshman@iiitdwd.ac.in)

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Perceiving randomness and probabilistic problem . . . . .	7
1.2	Statistics and randomness . . . . .	8
1.3	Strategies for solving probabilistic problems . . . . .	9
1.4	Overview . . . . .	10
<b>2</b>	<b>Elementary probability</b>	<b>12</b>
2.1	What is probability? . . . . .	12
2.2	Basic concepts of probabilistic modelling . . . . .	17
2.3	Types of probabilistic models . . . . .	29
2.4	Conditional probability . . . . .	37
<b>3</b>	<b>Probability distribution</b>	<b>51</b>
3.1	Random variable . . . . .	52
3.2	Discrete distribution . . . . .	53
3.3	Continuous distribution . . . . .	92
3.4	General random variable . . . . .	131
<b>4</b>	<b>Large sample theory</b>	<b>132</b>
4.1	Sample statistic . . . . .	133
4.2	Probabilistic inequalities . . . . .	137
4.3	Limit theorems . . . . .	147

<b>5 Statistical methods</b>	<b>155</b>
5.1 Point estimation . . . . .	155
5.2 Interval estimation . . . . .	184
5.3 Hypothesis testing . . . . .	190
<b>6 Statistical modelling</b>	<b>195</b>
6.1 Statistical models . . . . .	195
6.2 Linear models . . . . .	197
6.3 Generalized linear models . . . . .	215
6.4 Future note: control, discovery models, and generation .	218
<b>References</b>	<b>221</b>

## **Abstract**

This text presents an unified approach of probability and statistics in the pursuit of understanding and computation of randomness in engineering or physical or social system with prediction with generalizability. Starting from elementary probability and theory of distributions, the material progresses towards conceptual and advances in prediction and generalization in statistical models and large sample theory. We also pay special attention to unified derivation approach and one-shot proof of each and every probabilistic concept. Our presentation of intuitive and computation framework of conditional distribution and probability are strongly influenced by unified patterns of linear models for regression and for classification. The text ends with a future note on the unified approximation of the linear models, the generalized linear models and the discovery models to neural networks and a summarized ML system.

## **preface**

---

pothee padhi padhi jag mua, pandit bhaya na koy |  
dhaee aakhar prem ka, padhe so pandit hoy |  
Kabir.

Here, prem may refer, devotion to observation, understanding, & applicability.

The central objects of our studies are probability as a measure of randomness and statistics as a computation of a sample statistic (with goodness of fit) from given data-sets.

### **Objective and approach**

This text assumes that you know approximately nothing about probability and statistics. Its goal is to provide you with observation, understanding and applications of each every concepts that ultimately leads to solving real world problems.

We cover a large number of techniques, from the simplest and most commonly used (such as conditioning) to some of the advance techniques (e.g. estimation and generalization of a statistical model).

Computational framework (in the intersection of probability & statis-

tics, optimization & numerical methods, algorithm & programming, and computer) means a practical understanding of a concept and its applicability in solving a related problem in computers effectively with the help of some numerical method. The text favours a computational approach of each and every concept with a better observation, understanding and applicability through concrete working examples and their simulation in computer. Furthermore, the approach in order to solve a real world problem characterizes with an accuracy, (i.e. approximately very close to the correct/optimal result), an efficiency (i.e. a polynomial time computational and iteration complexities), a robustness (i.e. work for every possible realization of a random sample of any size) and a stability (i.e. not sensitive to small perturbations in the input).

**अभिकलनात्मक ढाँचा:** यह प्रायिकता एवं सांख्यिकी, इष्टतमीकरण एवं संख्यात्मक पद्धतियां, कलन विधि एवं कार्यक्रम निर्माण और संगणक के प्रतिच्छेदन में निहित है।

**अभिकलन :** पुं० [सं० अभिकल्(गिनना)+ल्युट-अन] [भ० क० अभिकलित] परिकलन का वह प्रकार या रूप जिसमें अनुभवों, बाहरी घटनाओं, निश्चित सिद्धान्तों आदि से भी सहायता की जाती है।

**Computation:** The type or form of computation in which experiences, external events, definite principles etc. are also helped.

**परिकलन :** पुं० [सं० परिकल् (गिनना)+णिच+ल्युट-अन] [भ० क० परिकलित] गणित में वह गणना जो कुछ जटिल होती है तथा जिसमें कुछ विशिष्ट तथा निश्चित क्रियाओं की सहायता लेती पड़ती है।

**Calculation:** The calculation which is somewhat complex and requires the help of some specific and definite actions.

**ढाँचा :** पुं० [सं० स्थाता] कोई रचना बनाते समय उसके विभिन्न अंगों-उपांगों का प्रयोग कर निर्मित वह आरंभिक या स्थूल रूप जिस पर बाकी सारी रचना प्रस्तुत होती है। जैसे-मकान का ढाँचा, कुरसी का ढाँचा।

**Framework:** While making a composition, the initial or gross form created by using its various parts and appendages on which the rest of the composition is presented (e.g., framework of a house, framework of a chair).

**Figure 1:** A self-explanatory meaning of computational framework in Hindi with corresponding translation in English.

## Prerequisite

This text assumes that you have some basic understanding of linear algebra, calculus, programming and algorithm.

## Roadmap

The central objects of our studies are probability as a measure of randomness and statistics as a computation of a sample statistic (with goodness of fit) from given data-sets. So, the text is decomposed into two parts.

Part I, probability, covers the following topics:

- i. What is probability? What is probabilistic modelling of a random experiment? What are the types of probabilistic modelling? How a prior information or conditioning affects a probabilistic modelling?
- ii. The central concept of probability, probability distribution, as a computational measure of randomness in a random experiment.
- iii. The most common probability distributions and their derived distributions.
- iv. Joint probability distributions and their derived distributions, convolutions.
- v. Central representation of a probability distribution: expectation, variance, co-variance.
- vi. Moment generating function and its applications

Part II, statistics, covers the following topics:

- i. What is sample? What is data from a sample? What is sample statistic? What is stochastic convergence?
- ii. Probabilistic inequalities and non-asymptotic large sample theory.

- iii. Limit theorems and asymptotic large sample theory.
- iv. Methods of point estimation and goodness of fit (e.g. bias, variance, efficiency, sufficiency).
- v. Interval estimation and limit theorems.
- vi. Hypothesis testing and statistical inferences
- vii. Linear models and prediction (e.g. simple univariate linear models, simple multivariate linear models, embedded linear models, multiple multivariate linear models).
- viii. Simulation with Python.

Note: don't jump into solving a noisy observation based problem in a computer too hastily: while solving the problem in computer is no doubt one of the most exciting areas in computational mathematics, you should master the fundamentals first.

### **Silent features of the book**

- a. Derivation of various concepts such as:
  - i. computation of probability of an subinterval  $[a, b]$  in the sample space  $\Omega$  of a continuous probability model
 
$$P([a, b]) = \frac{\text{length of } [a, b]}{\text{length of } \Omega} = \frac{b - a}{\text{length of } \Omega}.$$
  - ii. conditional probability as a technique to understand, compute and apply to central concepts such as concepts joint probability, total probability, Bayes rule, conditional distribution, joint distribution and conditional expectation.
- b. presentation of each and every example in an ordered 3-tuple of backdrop, question and answer.
- c. One-shot derivation of various proofs and solutions such as derived distribution, limit theorem, point estimation, interval estimation and computation & estimation of parameters for a linear model.

- d. A unified approach to compute and estimate parameters of simple univariate, simple multivariate, embedded and multiple multivariate linear models.
- e. Goodness of fit for every estimate of a parameter or a parametric model.
- f. Python demonstration of almost every probability distribution, large sample theory and statistical methods.
- g. An effort to utilise self-explanatory capability of a word in Indian languages to understand, observe and apply at the best (i.e. upto optimality) in-line with new education policy 2023.

<p><b>सांख्यिकी:</b>  <b>सांख्यिकःवि०[सं०]</b>  <b>सांख्य+इ(तच)+अक</b></p>	<p>किसी विषय की (यथा—उत्पादन, जन्मरण, रोग आदि की) संख्याएँ एकत्र करके उनके आधार पर कुछ सिद्धांत स्थिर करने या निष्कर्ष निकालने की विद्या। स्थिति-शास्त्र।</p>
<p>The science of collecting numbers on any subject (e.g., production, birth-death, diseases, etc.) and establishing some principles or drawing conclusions on their basis. situation</p>	

**Figure 2:** A self-explanatory meaning of statistics in Hindi with corresponding translation in English.

- h. This is a text, not a reference book, for probability and statistics of undergraduate studies. I have tried to present the fundamental ideas of the subject in simple and intuitive contexts in computational frameworks.

# 1

---

## Introduction

---

The purpose of the text is to present the computational framework of probability and statistics as a mathematical discipline that observe, understand and applicability of a problem based on probabilistic phenomenon or a mathematical model based on noisy measurements.

### 1.1 Perceiving randomness and probabilistic problem

Natural sciences (e.g. theory of errors in measurements, problems of the theory of ballistics, problems of statistics) made it necessary to develop the theoretical probability with advanced analytical tools.

The modern development of probability is characterized by the notions of measure theory and, subsequently, functional analysis that lead to an extended content of theoretical probability (e.g. large sample theory, stochastic processes, stochastic modelling).

In the middle of 20th century, applicability of probability developed in modern science immeasurably (e.g. molecular orbital theory, statistical physics, microscopic understanding of diffusion and transport of particles through interaction, intra-action and characterization of motion). At microscopic level, particle theory establish a fact that every sub-

stance is composed of very many particles i.e. in a homogeneous body their properties are very similar. Under these circumstances standard mathematical research methods are useless in the investigation of physical theories. For example, the techniques of differential equations are incapable of yielding any serious results under these conditions.

**Example 1.1.** Consider a diffusion model

$$\frac{\partial u}{\partial t} = D \frac{\partial^2}{\partial t^2} (u). \quad (1.1)$$

Then,

- a. At macroscopic level, the equation (1.1) refers a deterministic system.
- b. While, at microscopic level, the motion of each individual molecule/particle is having a Brownian motion, which is random in nature.

Schematic a phenomenon or selection of mathematical tools to investigate the phenomenon. Goodness of fit or quality can be judged by how well the theory agrees with experiment and with practice.

## 1.2 Statistics and randomness

Consider a statement, "80% chance of rain on this Friday". So, we perceive the following:

- a. the chance 80% is estimated from data statistically (i.e. a statistical estimate).
- b. an estimation is performed in a deterministic way, because it involves a computation technique.
- c. so, a statistical estimation is carried through measurement (or observation) using a computation technique.
- d. furthermore, a measurement involve measurement randomness due to computation approximation and measuring unit.

### 1.3 Strategies for solving probabilistic problems

The central objects of our studies are probability as a measure of randomness and statistics as a computation of a sample statistic (with goodness of fit) from given data-sets. So, the text is decomposed into two parts.

Part I, probability, covers the following topics:

- i. What is probability? What is probabilistic modelling of a random experiment? What are the types of probabilistic modelling? How a prior information or conditioning affects a probabilistic modelling?
- ii. The central concept of probability, probability distribution, as a computational measure of randomness in a random experiment.
- iii. The most common probability distributions and their derived distributions.
- iv. Joint probability distributions and their derived distributions, convolutions.
- v. Central representation of a probability distribution: expectation, variance, co-variance.
- vi. Moment generating function and its applications

Part II, statistics, covers the following topics:

- i. What is sample? What is data from a sample? What is sample statistic? What is stochastic convergence?
- ii. Probabilistic inequalities and non-asymptotic large sample theory.
- iii. Limit theorems and asymptotic large sample theory.
- iv. Methods of point estimation and goodness of fit (e.g. bias, variance, efficiency, sufficiency).
- v. Interval estimation and limit theorems.

- vi. Hypothesis testing and statistical inferences
- vii. Linear models and prediction (e.g. simple univariate linear models, simple multivariate linear models, embedded linear models, multiple multivariate linear models).
- viii. Simulation with Python.

Note: don't jump into solving a noisy observation based problem in a computer too hastily: while solving the problem in computer is no doubt one of the most exciting areas in computational mathematics, you should master the fundamentals first.

## **1.4 Overview**

- a. Derivation of various concepts such as:
  - i. computation of probability of an subinterval  $[a, b]$  in the sample space  $\Omega$  of a continuous probability model
 
$$P([a, b]) = \frac{\text{length of } [a, b]}{\text{length of } \Omega} = \frac{b - a}{\text{length of } \Omega}.$$
  - ii. conditional probability as a technique to understand, compute and apply to central concepts such as concepts joint probability, total probability, Bayes rule, conditional distribution, joint distribution and conditional expectation.
- b. presentation of each and every example in an ordered 3-tuple of backdrop, question and answer.
- c. One-shot derivation of various proofs and solutions such as derived distribution, limit theorem, point estimation, interval estimation and computation & estimation of parameters for a linear model.
- d. A unified approach to compute and estimate parameters of simple univariate, simple multivariate, embedded and multiple multivariate linear models.

- e. Goodness of fit for every estimate of a parameter or a parametric model.
- f. Python demonstration of almost every probability distribution, large sample theory and statistical methods.
- g. An effort to utilise self-explanatory capability of a word in Indian languages to understand, observe and apply at the best (i.e. upto optimality) in-line with new education policy 2023.

<p><b>सांख्यिकी:</b>  <b>सांख्यिकःवि०[सं०]</b>  <b>सांख्य+इ(तच)+अक</b></p>	<p>किसी विषय की (यथा—उत्पादन, जन्मरण, रोग आदि की) संख्याएँ एकत्र करके उनके आधार पर कुछ सिद्धांत स्थिर करने या निष्कर्ष निकालने की विद्या। स्थिति-शास्त्र।</p>
<p>The science of collecting numbers on any subject (e.g., production, birth-death, diseases, etc.) and establishing some principles or drawing conclusions on their basis. situation</p>	

**Figure 1.1:** A self-explanatory meaning of statistics in Hindi with corresponding translation in English.

- h. This is a text, not a reference book, for probability and statistics of undergraduate studies. I have tried to present the fundamental ideas of the subject in simple and intuitive contexts in computational frameworks.

# 2

---

## **Elementary probability**

---

### **2.1 What is probability?**

Probability is a mathematical discipline that concerning with numerical descriptions of how likely an event is to occur or a degree of belief of the occurrence of an event, or how likely it is that a proposition is true. The probability of an event is a number between 0 and 1. In simple, probability is the measure or quantification of randomness that falls in the intersection of mathematical thinking and intuitive observation/experience of patterns. It finds applications in . It diffuses through almost every discipline of sciences & mathematics, engineering & technology and social & life sciences. And it finds applications in inference, decision making, experimental design, reliability theory, statistical theory of information, communication, control, and many more except motion of planets. A few example inspired from [Pfeiffer and Schum, 1973] and [Ramon van Handel, 2006]:

- a. Physics: a few quantities (e.g. temperature and pressure) arise as a direct consequence of the random motion of atoms and molecules. Heisenberg uncertainty principle in quantum mechanics.

- b. Biology and medicine: in human evaluation, random mutations lead to the amazing diversity of life. Stochastic models are essential in understanding the spread of disease, both in a population (e.g. epidemics) or in the human body (e.g. cancer).
- c. Chemistry: a chemical reaction takes place when molecules randomly meet. Stochastic models of chemical kinetics are particularly important in systems with very low concentrations (e.g. biochemical reactions in a single cell).
- d. Electronics and communication engineering: the effect of noise to design a reliable communication protocol that we use on a daily basis in our cell phones. Observation or measurement of a signal using a suitable measuring unit happens to be noisy in nature (i.e. in practice, measurement of a signal contains a signal along-with an additive noise).
- e. Computer science: randomness in an algorithm (for some best known method) brings robustness and stability in order to compute a solution of a complex problem.
- f. Finance and economics: stock and bond prices are inherently unpredictable, so these lead to some kind of stochastic models. The modelling of randomly occurring rare events forms the basis for all insurance policies, and for risk management in banks.
- g. Sociology: stochastic models provide basic understanding of the formation of social networks and of the nature of voting & polling schemes.
- h. Statistics and machine learning: statistical methods lead to stochastic models that form the foundation for almost all of data science (e.g. tabular data science, image data science, text data science or NLP).
- i. Optimization: in order to come out of stuck region of local in minimization a function, there is a need noise in an optimization algorithm through noisy or perturbed gradient (e.g. noisy gradient descent, stochastic gradient descent).

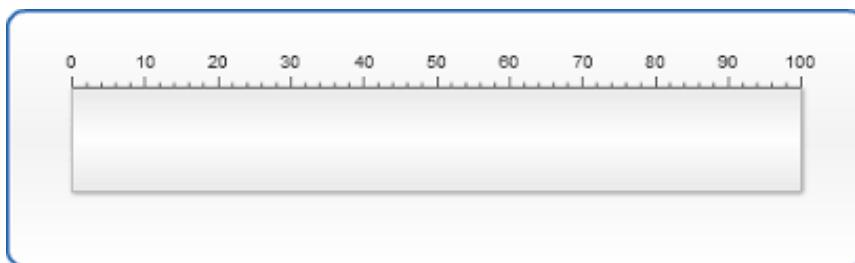
**प्रायिकता:** प्रायिकता एक गणितीय अध्ययन का विषय है जो किसी घटना के घटित होने की कितनी प्रायिकता या विश्वास की मात्रा के संख्यात्मक वर्णन से संबंधित है।

**प्रायिक:** विं [सं० प्राय+ठक्-इक] [आव० प्रायिकता] जो नियमित रूप से या सदा तो नहीं फिर भी बीच-बीत में प्रायः होता रहता हो।

Which does not happen regularly or always, but still happens often from time to time.

**Figure 2.1:** A self-explanatory meaning of probability in Hindi with corresponding translation in English.

**Example 2.1** (A measurement error due to limitation of a scale). Consider a measurement scale (2.2) that can measure 0 to 100 cm with 1cm accuracy. Now if we have to measure 99.73cm with respect to the given scale, we find that 99.73 cannot be measured deterministically on it. With respect to the scale, 99.73 has three significant figures, so it is  $99.73 \approx 99.7$ , where first two numbers i.e. 99 are deterministic significant figures and the third one .7 is a uncertain or estimated or random significant figure. So, we can a measurement includes uncertainty due to limitation of the corresponding measuring instrument in use.

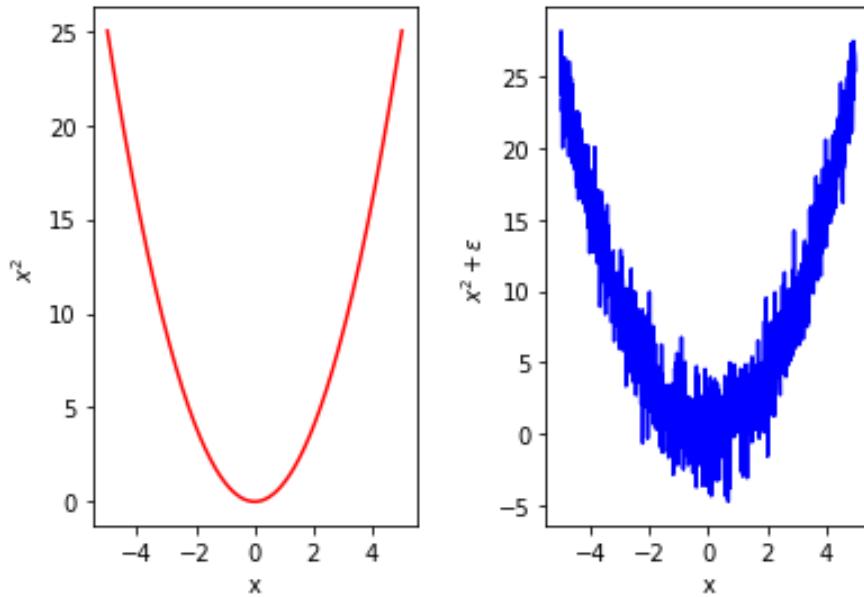


**Figure 2.2:** a measurement scale that can measure 0 to 100 cm with 1cm accuracy.

**Example 2.2** (the toss of a fair coin). You know intuitively that there is a 50 percent chance of getting heads, and 50 percent chance of getting

tails. If you want to do the math to calculate the probability of a head, here's the basic formula that counts the number of times that the event will happen (e.g. in this case, there's just one chance of a head appearing, so it's 1). Divide this by the total number of possible outcomes (e.g. with a coin, it's either heads or tails, which is 2 possible outcomes). So the probability of getting a head is 0.5 or 50 percent.

**Example 2.3** (Estimating a number or a function). Consider an estimation of a number from its noisy measurement. For example,  $\bar{x} = \mu + \varepsilon$ , where  $\mu$  is the true mean, and  $\bar{x}$  is the estimated mean with a noise  $\varepsilon$  as the error containing some randomness. Furthermore, consider an estimation of a deterministic function of  $f(x) = x^2$  with its noisy measurement  $x^2 + \varepsilon$  in figure (2.3).



**Figure 2.3:** In left there a deterministic function  $x^2$  to be estimated from its noisy version in right.

**Example 2.4.** Backdrop: three men are on a trip. One of the three will volunteer to cook breakfast. A is twice as likely to volunteer as B, and B is three times as likely to volunteer as C.

Question: compute the probability of volunteering for each man.

Answer: suppose that one and only one will volunteer among A, B, and C. So, we can assume that  $\{A, B, C\}$  are mutually exclusive and totally exhaustive, i.e.  $P(A) + P(B) + P(C) = 1$ .

Now, from the question, we have the following assumptions:

$$(a.) P(A) = 2P(B), \quad (b.) P(B) = 3P(C), \quad (c.) P(A) + P(B) + P(C) = 1 \\ \implies 2 \cdot 3P(C) + 3P(C) + P(C) = 1 \implies P(C) = \frac{1}{10}.$$

And hence, we get

$$P(A) = \frac{6}{10}, \quad P(B) = \frac{3}{10}, \quad \text{and} \quad P(C) = \frac{1}{10}.$$

**Remark 2.1.** In summary, we may perceive following at first instant observation from probability:

- a. chance/unpredictable/random/stochastic,
- b. possible outcomes in a random experiment (an experiment with unpredictable possible outcome in each of its trials).

So, generally we come across a random experiment with a list of possible outcomes, but a possible outcome in each trial of the experiment is unpredictable (i.e. random)

**Definition 2.1 (Random experiment).** A random experiment is an experiment whose outcome is not predictable before the experiment is performed, however in advance we know all possible outcomes in the experiment. That is, an outcome of a random experiment is uncertain and a probabilistic modelling is providing an approach to the quantify the uncertainty.

Furthermore, there is no restriction on what constitutes an experiment (e.g. it could be a single toss of a coin, or three tosses, or an infinite sequence of tosses). So, it is important to note that in formulation of a probabilistic model, there is only one experiment (e.g. three tosses of a coin constitute a single experiment, rather than three experiments).

**Example 2.5** (Tossing a coin). If we toss a coin, then either a head or tail comes up with some degree of belief, but we don't know in advance which one will turn up.

**Example 2.6** (Tossing a coin). In weather prediction, a meteorologist just know either it will rain or not rain with degree of belief, but we doesn't know in advance which one will turn up.

## 2.2 Basic concepts of probabilistic modelling

Consider a few situation—(a.) the probability that it will rain the next day or (b.) the probability that you will win the lottery [Kay, 2006]. Here, we cannot be certain because, there are many factors that affect the weather and those leads to uncertainty that whether it will or not rain the next day. So, as a predictive tool we usually assign a number between 0 and 1 as our degree of certainty or belief that the event, rain, will occur. If we say that there is a 30% chance of rain, means we believe that if identical conditions prevail, then 3 times out of 10, rain will occur the next day. To predict at the nature's standard, it is essential to come up with an accurate model by knowing everything of science of meteorology that gives the probability is either 0 or 1. Unfortunately, it is not possible at practice, i.e. uncertainty of an outcome is not because of the inaccuracy of our model, but because the experiment has been designed to produce uncertain results.

In the last section (2.1), we see a list of common items e.g. a random experiment, a set of possible outcomes, and the probabilities assigned to these outcomes. These attributes are common to all probabilistic descriptions. And hence, it is very much essential to come up with a systematic approach as designed by Kolmogrow to compute randomness or uncertainty through three basic concepts of probabilistic modelling of random experiment:

- (a.) a sample space,
- (b.) an event of interest, and
- (c.) probability measure of an event as an assignment or degree of belief between 0 and 1.

### 2.2.1 A sample space i.e. specification of all possible outcomes

Once a random experiment is being performed, we need to specify all possible outcomes of the random experiment as:

- (a.) First we identify an outcome of the random experiment,
- (b.) Secondly, we list out all possible outcomes of the experiment as a sample space.

**Example 2.7** (Throwing two dice together). Question: Throw two dice together and find its sample space.

Answer: Consider a throw of two dice together, then

- a. we identify an outcome as an ordered pair  $(i, j)$ , where the first die comes up  $i$  and the second die comes up  $j$ .
- b. we list out all possible outcomes as  $\{(i, j) | 1 \leq i \leq 6, 1 \leq j \leq 6\}$ .

So, we have the sample space,

$$\Omega = \{(i, j) | 1 \leq i \leq 6, 1 \leq j \leq 6\} \text{ with cardinality } |\Omega| = 36.$$

**Example 2.8** (Waiting time for a bus at a bus stop). Question: suppose you are waiting for a bus at a bus stop, then compute its sample space.

Answer: If you are waiting for a bus at a bus stop, then

- a. a real number  $t \geq 0$  is identified as a possible outcome of the experiment.
- b.  $t = 0h$  refers an outcome that on your arrival at the bus stop, immediately you get a bus.  $t = 1.5h$  refers an outcome that on your arrival at the bus, you waited for  $1.5h$  to get a bus. In this process, listing out all possible outcomes as  $\{t \in \mathbb{R} | t \geq 0\}$ .

So, the sample space is given by

$$\Omega = \{t \in \mathbb{R} | t \geq 0\} = [0, \infty).$$

**Example 2.9** (Path of a bee for 5 seconds). Question: consider a random experiment of observing path of a bee for 5 seconds. Compute its sample space.

Answer: If we observe path of a bee for 5 seconds, then

- a. a possible outcome is identified as a continuous map  $\omega : [0, 5] \rightarrow \mathbb{R}^3$  in (2.4):



**Figure 2.4:** Path of a bee as continuous map  $\omega : [0, 5] \rightarrow \mathbb{R}^3$ .

- b. listing all possible outcomes as

$$\Omega = \{\omega : [0, 5] \rightarrow \mathbb{R}^3 | \omega \text{ is a continuous map}\}.$$

So, the sample space is given by

$$\Omega = \{\omega : [0, 5] \rightarrow \mathbb{R}^3 | \omega \text{ is a continuous map}\}.$$

### 2.2.2 The event (of interest)

**Definition 2.2 (Informal).** Once we have identified a sample space for a given random experiment, then we can come up with a statement in order to construct (or define) an event as a collection of some of outcomes (or favourable outcomes) satisfying the statement.

**Definition 2.3 (Formal).** A subset  $A$  of the sample space  $\Omega$  which satisfies a given statement i.e. an element of a  $\sigma$ -algebra of the sample space  $\Omega$  (i.e.  $A \in \sigma(\Omega)$ ). An  $\sigma$ -algebra of  $\Omega$  is collection of subsets in  $\Omega$  and closed with respect to complementation and countable union of its members.

An event of interest can be computed as follows:

- First, list out all possible outcomes as a sample space of the given random experiment.
- Secondly, collect some possible outcomes satisfying a statement under interest/consideration.

**Example 2.10.** Question: consider an experiment of throwing a die and getting a number exactly divisible by 3.

Answer:

- Listing out all possible outcomes as

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

- $A \rightarrow$  collection of some possible outcomes satisfying a statement that a number is exactly divisible by 3. So, we have

$$A = \{3, 6\}.$$

**Example 2.11 (A throw of two dice).** Question: consider a throw of two dice together. Getting an outcome with a sum of is 7.

Answer:

- When we toss two coins together, the sample space is given by

$$\Omega = \{(i, j) | i, j = 1, 2, 3, 4, 5, 6\}.$$

- $A \rightarrow$  collection of some possible outcomes satisfying a statement that sum is 7. So, we have

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.$$

**Example 2.12 (Waiting time for a bus at a bus stop).** Question: Suppose you are waiting for bus at a bus stop. Getting a bus in the first hour.  
Answer:

- As you are waiting for a bus at a bus stop, so, the sample space is given by

$$\Omega = \{t \in \mathbb{R} | t \geq 0\} = [0, \infty).$$

- b.  $A \rightarrow$  collection of some possible outcomes satisfying a statement that a bus is in first. So, we have

$$A = \{t \in \mathbb{R} | 0 \leq t \leq 1\} = [0, 1] \subset \Omega.$$

**Example 2.13** (Path of a bee for 5 seconds). Question: In a random experiment of observing path of a bee for 5 seconds. Observing path of bee stuck in a region C for first 1 second.

Answer:

- a. If we observe a bee for 5 second, so, the sample space is given by

$$\Omega = \{\omega | \omega : [0, 5] \rightarrow \mathbb{R}^3 \text{ is a continuous map}\}.$$

- b.  $A \rightarrow$  collection of some outcomes satisfying a statement that a bee got stuck in a region C in first second. So, we have

$$A = \{\omega : [0, 5] \rightarrow \mathbb{R}^3 | \omega(t) \in C \text{ for } t \in [0, 1]\} \subset \Omega.$$

**Remark 2.2.** a. If  $\omega \in \Omega$  is an outcome of a random experiment, then an event  $A$  occurs if  $\omega \in A$ .

- b.  $A^c$  occurs if  $A$  does not occur.  
 c.  $A \cup B$  occurs if A occurs or B occurs (or both).  
 d.  $A \cap B$  occurs if A occurs and B occurs.  
 e.  $A \setminus B = A - B = A \cap B^c$  occurs if A occurs and B does not occur.  
 f.  $A$  and  $B$  are mutually disjoint if  $A \cap B = \varnothing$  i.e. A occurs and B does not occurs

### 2.2.3 Probability Measure

Till now, we have seen the following:

- (a.) Sample space ( $\Omega \rightarrow$ ) as a collection of all possible outcomes from the random experiment.
- (b.) An event ( $A$ ) defined by a statement as a collection of some possible outcomes from the experiment.

**Definition 2.4** (Probability measure of an event). Now, we define probability measure of an event A as a degree of confidence or a map  $P : \sigma(\Omega) \rightarrow [0, 1]$  satisfies the following properties or axioms:

- (i.)  $P(\Omega) = 1, P(\varphi) = 0,$
- (ii.)  $P(A^c) = 1 - P(A) \in [0, 1]$  for any event  $A \in \sigma(\Omega),$
- (iii.)  $P(A \cup B) = P(A) + P(B)$  for any two disjoint events A and B  
i.e.  $A \cap B = \varphi,$

The third axiom or property can be generalized to ( $\sigma$ - additive of) any countable collection of mutually disjoint events  $A_i$  as follows:

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \text{ for any mutually disjoint events i.e.}$$

$$A_i \cap A_j = \varphi \text{ for } i \neq j.$$

A probability measure of an event A can be computed using a suitable probability law (e.g. a uniform law or symmetrical outcomes) in the following two steps:

- (01) first come up with an assumption or information from the random experiment, where first instant assumption happens to be uniform assumption (or equally likely outcomes or symmetrical outcomes). For example, in a single toss of a coin, the chances of getting both head and tail are equal due to symmetrical coin (with respect to head and tail).
- (02) then computing the probability measure of an event using the assumption observed from the random experiment in step-01 and the axioms (2.4) of the probability measure of the event.

**Example 2.14** (A roll of a die). Question: In a roll of a fair die, compute probability of occurrence of each and every outcomes.

Answer: In a roll of a fair die, the sample space is given by

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

As the die happens to be fair, so each outcome is equally likely in the roll of the fair die i.e. there is a uniform law of distribution of outcomes. So, we have

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) \quad (2.1)$$

Now, using axioms of the probability measure for an event, we have

$$\begin{aligned} P(\Omega) &= 1, \Omega = \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\} \implies \\ P(1) + P(2) + P(3) + P(4) + P(5) + P(6) &= 1 \end{aligned} \quad (2.2)$$

Solving for equations (2.1) and (2.2), we get

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.$$

**Example 2.15** (Swimming in leisure time per week a day). Ram likes swimming in leisure time per week on a day. An outcome is a day that is

$$\Omega = \{\text{Monday}, \text{Tuesday}, \text{Wednesday}, \text{Thursday}, \text{Friday}, \text{Saturday}, \text{Sunday}\}.$$

The first instant assumption is the uniform as each day is equally likely, so we have

$$\begin{aligned} P(\text{Monday}) &= P(\text{Tuesday}) = P(\text{Wednesday}) = P(\text{Thursday}) \\ &= P(\text{Friday}) = P(\text{Saturday}) = P(\text{Sunday}). \end{aligned}$$

Now, using axioms of probability measure of an event, we have

$$\begin{aligned} P(\text{Monday}) + P(\text{Tuesday}) + P(\text{Wednesday}) + P(\text{Thursday}) \\ + P(\text{Friday}) + P(\text{Saturday}) + P(\text{Sunday}) &= 1. \end{aligned}$$

By solving above two equations, we get

$$\begin{aligned} P(\text{Monday}) &= P(\text{Tuesday}) = P(\text{Wednesday}) = P(\text{Thursday}) \\ &= P(\text{Friday}) = P(\text{Saturday}) = P(\text{Sunday}) = 1/6. \end{aligned}$$

**Example 2.16** (swimming on weekday as often as weekend). Ram goes for swimming on weekday as often as weekend. The first instant assumption says that weekdays and weekends are equally likely. So

$$P(\text{Weekdays}) = P(\text{Weekends}) = 1/2.$$

Further, there are 5 days in weekdays, and each of these 5 days are equally likely, so, we have

$$\begin{aligned} P(\text{Monday}) &= P(\text{Tuesday}) = P(\text{Wednesday}) = P(\text{Thursday}) \\ &= P(\text{Friday}) = 1/2 \cdot 1/5 = 1/10. \end{aligned}$$

And there are two days in weekends, and each of these 2 days are equally likely, so, we have

$$P(\text{Saturday}) = P(\text{Sunday}) = 1/2 \cdot 1/2 = 1/4$$

**Example 2.17** (Continuation of information from the previous two examples). He goes on Friday equals to Sum of Wednesday and Thursday. Let  $P(\text{Wednesday}) = P(\text{Thursday}) = p$ . Then  $P(\text{Friday}) = 2p$  and  $P(\text{Monday}) = P(\text{Tuesday}) = \frac{1}{2} \cdot \left(\frac{1}{2} - 4p\right) = \frac{1}{4} - 2p > 0 \implies 0 < p < 1/8$ .

Note: probability measure is not assigned to an individual outcome, it is assigned to an event containing the outcome.

**Example 2.18.** Five balls numbered 1, 2, 3, 4, 5 are drawn from an urn without replacement. Then compute the probability that they will be drawn in the same order as their number.

Each outcome is represented by the 5-tuple  $(z_1, z_2, z_3, z_4, z_5)$ , so total possible outcomes are  $|\Omega| = 5! = 120$ . The only outcome in the event  $A$  is  $(1, 2, 3, 4, 5)$ . Hence, the desired probability is

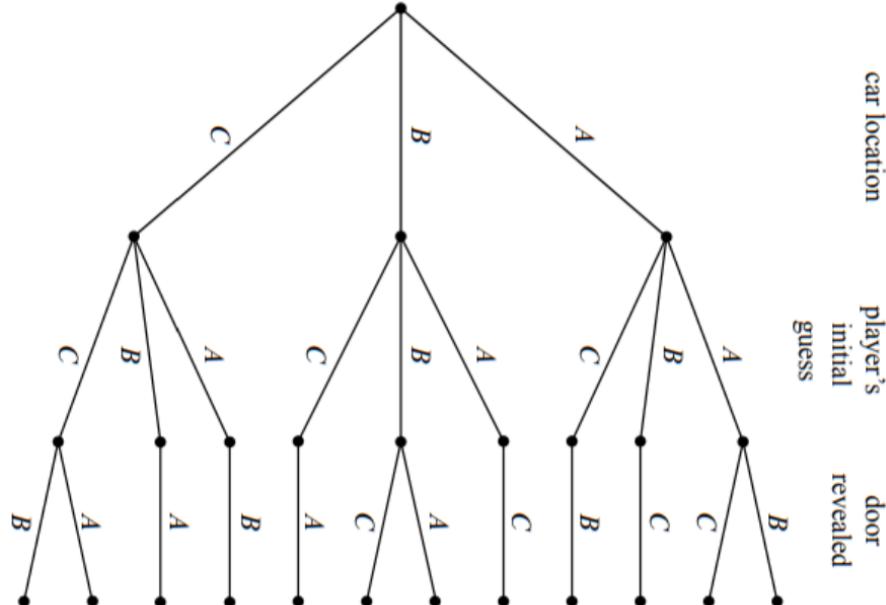
$$P(A) = \frac{1}{120}.$$

**Example 2.19** (Monte Hall problem). Background: suppose you're on a game show, and you're given the choice of three doors: behind one door there is a car, behind the others, there are goats. You pick a door (say No. 1), and the host, who knows what's behind the doors, opens another door (say No. 3), which has a goat. He then says to you, 'do you want to pick door No. 2?'

Question: Compute the probability with which the player wins by switching.

Answer: first we need to all possible outcomes followed by an event of interest (i.e. the player wins by switching) and probability measure of the event.

- a. In order to identify an outcome, we recall a randomly determined quantities that may help to determine the outcome as:
  - i. the door concealing the car (as a first layer of information),
  - ii. the door initially chosen by the player (as a second layer of information),
  - iii. the door that the host opens to reveal a goat (as a third layer of information).



**Figure 2.5:** Three randomly determined quantities are defining three layers of information upside down in the tree plot.

Consider the doors are A, B and C instead of 1, 2, and 3. Now as the first layer of the tree (see in figure (2.5)) is the door concealing

the prize. So, the prize may be behind any of the three doors A,B and C in equally likely manner.

The player could initially choose any of the three doors is the second layer of the tree (see in figure (2.5)).

Then a third layer of the tree (see in figure (2.5)) is that the host opens a door to reveal a goat.

If the prize is behind door A and the player picks door A, then the host could open either door B or door C. So, the possible outcomes are AAB or AAC.

If the prize is behind door A and the player picks door B, then the host must open door C. So, a possible outcomes is ABC.

Repeating all the possibility of above three randomly-determined quantities, then the leafs of the tree (see in figure (2.5)) determine all the 12 possible outcomes of the random experiment. So, sample space is given by

$$\Omega = \{AAB, AAC, ABC, ACB, BAC, BBA, BBC, BCA, CAB, CBA, CCA, CCB\}.$$

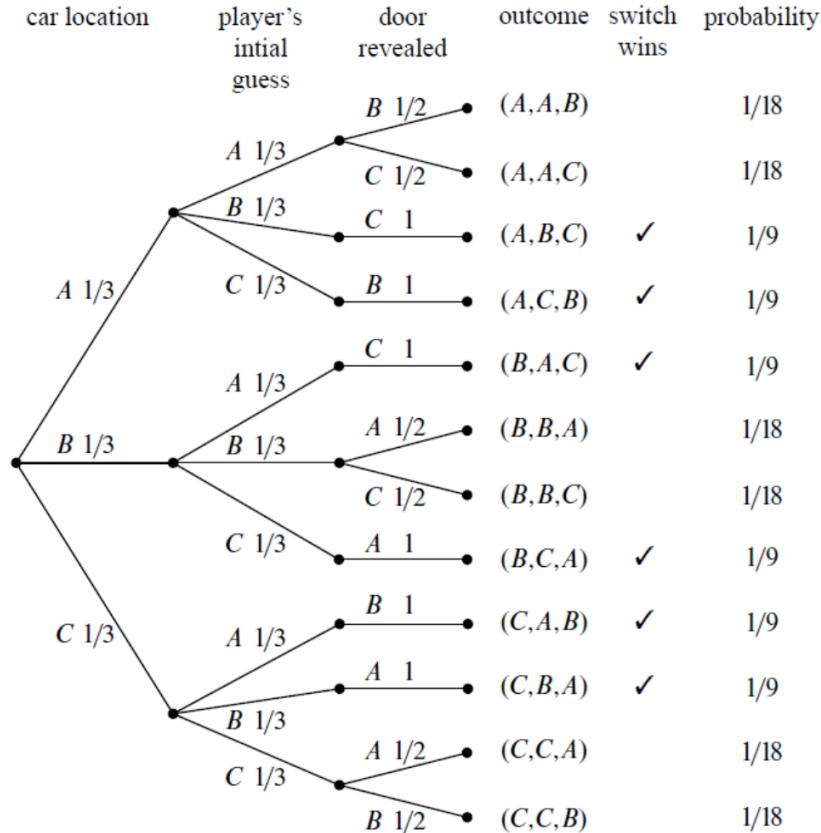
- b. The event of interest is determined by the statement that the player wins by switching. So, the desired event is given by

$$E = \{ABC, ACB, BAC, BCA, CAB, CBA\}.$$

- c. Computing probability measure of the desired event E. Consider the topmost event of the tree (see in figure (2.5)) i.e. the outcome AAB, then, the probability of AAB computed as:

$$P(AAB) = P(A)P(A|A)P(B|A, A) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}.$$

Similarly, we can compute probability of every possible outcome as a leaf of the tree (see in figure (2.5)) in the following probability tree diagram:



**Figure 2.6:** A plot of probability three is containing probability of every possible outcome of the Monte-Hall problem.

So, the desired probability is given by

$$P(E) = P(\{ABC, ACB, BAC, BCA, CAB, CBA\}) = 6 \cdot \frac{1}{9} = \frac{2}{3}.$$

**Example 2.20.** [sampling with order and replacement] Compute number of distinct non-negative integer-valued solutions of the equation

$$x_1 + x_2 + \dots + x_k = n. \quad (2.3)$$

Consider a sequence of  $n *$ 's and  $m-1 |$ 's. There is a bijection between such sequences and non-negative integer-valued solutions to the equation. For example, if  $m = 4$  and  $n = 3$ , i.e. a sequence of  $n = 3 *$ 's and

$m - 1 = 4 - 1 = 3$  |'s

$$\underbrace{\star \star}_{x_1=2} \mid \underbrace{\phantom{\star \star}}_{x_2=0} \mid \underbrace{\star}_{x_3=1} \mid \underbrace{\phantom{\star}}_{x_4=0}.$$

There are  $\binom{n+m-1}{n}$  sequences of  $n$  's and  $m-1$  |'s and, hence, the same number of solutions to the equation (2.3).

**Example 2.21.** Question: consider a urn with 15 identical (in shapes), but different (in colour) balls (i.e. an instant of seven white, two green, and six red). If you draw a ball from the urn, then compute the probability of drawing the ball from the urn.

Solution: Drawing a from the urn means either drawing a white as a type A, or drawing a red ball as a type B or drawing a green ball as a type C. So, out of the fifteen identical balls, A contains seven equally likely favourable outcomes, B contains six equally likely favourable outcomes and C contains two equally likely favourable outcomes. And hence, we have the following probabilities:

$$P(A) = \frac{7}{15}, \quad P(B) = \frac{6}{15}, \quad \text{and} \quad P(C) = \frac{2}{15}.$$

**Remark 2.3.** a. Two events  $A$  and  $B$  are mutually exclusive if  $P(A \cap B) = 0$ .

b.

**Theorem 2.1.** Any two mutually disjoint events are mutually exclusive, but converse is not true.

*Proof.* Consider  $A$  and  $B$  are two mutually disjoint events. Then, we have

$$A \cap B = \varphi \implies P(A \cap B) = P(\varphi) = 0.$$

So,  $A$  and  $B$  are two mutually exclusive.  $\square$

## 2.3 Types of probabilistic models

Consider the number of persons at a business location that are talking on their respective phones any-time between 9 : 00 AM and 9 : 10 AM. Clearly, the possible outcomes are  $0, 1, 2, \dots, n$ , where  $n$  is the number of persons in the office. On the other hand, if we are interested in the length of time a particular caller is on the phone during that time period, then the outcomes may be anywhere from 0 to  $T$  minutes, where  $T = 10$ . Now the outcomes are infinite in number since they lie within the interval  $[0, T]$ . In the first case, since the outcomes are discrete (and finite), we can assign probabilities to the outcomes  $\{0, 1, 2, \dots, n\}$ . An equi-probable assignment would be to assign each outcome a probability of  $1/(n + 1)$ . In the second case, the outcomes are continuous (and therefore infinite) and so it is not possible to assign a non-zero probability to each outcome. In the continuous case, it is not appropriate to ask for the probability that  $T$  will be exactly, e.g. 5 minutes, because this probability will be zero. Instead, we inquire as to the probability that  $T$  will be between 5 and 6 minutes.

So, based on discrete and continuum patterns of outcomes in a sample space of a random experiment, there are two types of probabilistic models:

- a. Discrete probabilistic model
- b. Continuous probabilistic model

### 2.3.1 Discrete probabilistic model

A probabilistic model of a random experiment is discrete if the sample space of the random experiment is finite or at most countable (i.e. there are finitely many outcomes or infinitely many outcomes which can be arranged into a simple sequence) i.e.  $\Omega = \{\omega\}_{k \in \mathbb{N}}$ . And a probability measure of an event  $A$  can be computed in the following two steps:

- a. Come up with an assumption or information from the random experiment, where first instant assumption happens to be uniform assumption (or equally likely outcomes).

- b. Computing the probability measure of an event using the assumption observed from the random experiment in step-01 and the axioms of the probability measure of the event.

**Theorem 2.2** (Discrete probabilistic modelling). A probabilistic modelling of a random experiment is discrete if

- a. the sample space,  $\Omega$ , of the experiment is discrete i.e.  $\Omega$  is at most countable i.e.  $\Omega$  contains at most a countable collection of possible outcomes of the experiment i.e.  $\Omega = \{\omega_n | \omega_n \text{ is a possible outcome}\}$ .
- b. an even,  $A$ , is at most countable as a sub-sequence of some possible outcomes i.e.  $\{\omega_{n_k} | \omega_{n_k} \text{ satisfies a given statement}\}$ .
- c. Probability measure of an event (from a random experiment with a finite sample space) is defined as the ratio of number outcomes occur in the event  $A$  to number outcomes occur in the sample space  $\Omega$ .

Note: If the sample space of a random experiment is a finite i.e.  $\Omega = \{x_1, x_2, \dots, x_n\}$ , then computation of probability measure of an event  $A$  is given by

$$P(A) = \frac{\text{number of outcomes occur in } A}{\text{number of outcomes occur in } \Omega}.$$

**Example 2.22** (Throwing a die). Consider a throw of a die, then  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Define an event  $A \rightarrow$  even face or outcomes in the experiment, then  $A = \{2, 4, 6\}$ . So,  $P(A) = 3/6 = 1/2$ .

**Example 2.23** (Tossing a coin twice). Consider two tosses of a coin, then  $\Omega = \{HH, HT, TH, TT\}$ . Define an event  $A \rightarrow$  at least one head, then  $A = \{HH, HT, TH\}$ . So,  $P(A) = 3/4$ .

**Example 2.24** (Drawing ball from an urn [Feller, 1968]). An urn has  $k$  red balls and  $n - k$  black balls. If two balls are chosen in succession and at random with replacement, then compute the probability of a red ball followed by a black ball.

To solve the problem, we first label the  $k$  red balls with  $1, 2, \dots, k$  and the black balls with  $k+1, k+2, \dots, n$ . In doing so the possible outcomes of the experiment can be represented by a 2-tuple  $(z_1, z_2)$ :  $z_1 = 1, 2, \dots, n, z_2 = 1, 2, \dots, n$ . A successful outcome is a red ball followed by a black one so that the successful event is  $A = \{(z_1, z_2) : z_1 = 1, 2, \dots, k, z_2 = k+1, k+2, \dots, n\}$ . Then the desired probability is given by

$$\begin{aligned} a. P(A) &= \frac{k(n-k)}{n^2} = \frac{k}{n} \left(1 - \frac{1}{n}\right) \text{ with replacement,} \\ b. P(A) &= \frac{k(n-k)}{n(n-1)} = \frac{k}{n} \left(1 - \frac{1}{n}\right) \frac{n}{n-1} \text{ without replacement.} \end{aligned}$$

**Example 2.25.** Question: if two fair dice are tossed, find the probability that the same number will be observed on each one. Next, find the probability that different numbers will be observed.

Solution: if we toss two dice together than the sample is given by

$$\Omega = \{(i, j) | 1 \leq i \leq 6, 1 \leq j \leq 6\} \implies |\Omega| = 36.$$

Now, probability that the same number will be observed on each one is given by

$$P(\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}) = 6 \cdot \frac{1}{36} = \frac{1}{6}.$$

And, the probability that different numbers will be observed is given by

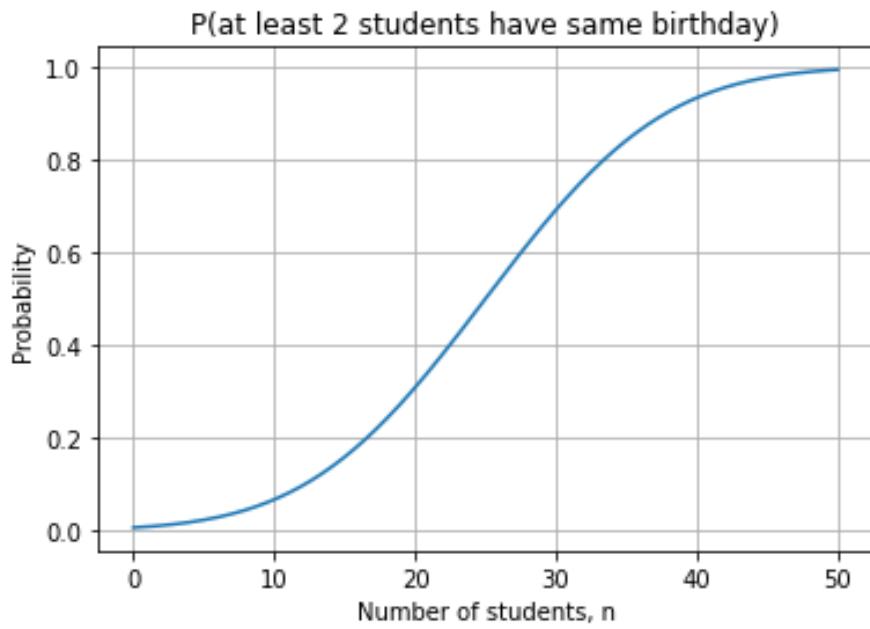
$$\begin{aligned} &P(\text{different numbers will be observed}) \\ &= 1 - P(\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}) \\ &= 1 - 6 \cdot \frac{1}{36} = 1 - \frac{1}{6} = \frac{5}{6}. \end{aligned}$$

**Example 2.26** (Birthday problem [Kay, 2006]). A probability class has  $n$  students enrolled. Then compute the probability that at least two of the students will have the same birthday.

We first assume that each student in the class is equally likely to be born on any day of the year. To solve this problem consider a birthday urn that contains 365 balls. Each ball is labelled with a different day

of the year. Now allow each student to select a ball at random, note its date, and return it to the urn. The day of the year on the ball becomes his/her birthday. The probability desired is of the event that two or more students choose the same ball. It is more convenient to determine the probability of the complement event or that no two students have the same birthday. Then, the desired probability is given by

$$\begin{aligned}
 & P(\text{at least 2 students have same birthday}) \\
 &= 1 - P(\text{no students have same birthday}) \\
 &= 1 - \frac{\binom{365}{n}}{365^n}.
 \end{aligned}$$



**Figure 2.7:** Probability of at least two students having the same birthday.

### 2.3.2 Continuous probabilistic model

A probabilistic model of a random experiment is continuous if the sample space of the random experiment is at least continuum in nature i.e.  $\Omega \neq \{x_k\}_{k \in \mathbb{N}}$ , but  $\Omega$  is an interval or union of intervals. In this case, a probability measure of an event  $A$  can be computed in the following two steps:

- a. Come up with an assumption or information from the random experiment, where first instant assumption happens to be uniform assumption (or equally likely outcomes).
- b. Computing the probability measure of an event using the assumption observed from the random experiment in step-01 and the axioms of the probability measure of the event.

**Remark 2.4 (Uniform law).** Consider a sample space  $\Omega = [a, b]$  with a uniform partition  $\{x_0 = a, x_1, x_2, \dots, x_n = b\}$  of width  $\delta = x_i - x_{i-1} = \frac{b-a}{n}$  i.e.  $[a, b] = [x_0, x_1] \cup [x_1, x_2] \cup \dots \cup [x_{n-1}, x_n]$  and the  $n$  sub-intervals are mutually disjoint except the common terminal points.

As per uniform law or equally likely assumption we have

$$\begin{aligned} P([x_0, x_1]) &= P([x_1, x_2]) = \dots = P([x_{n-1}, x_n]) \text{ and } P([x_{i-1}, x_i]) \propto \delta \\ \implies P([x_{i-1}, x_i]) &= c_i \delta \text{ for } i = 1, 2, \dots, n \implies c_1 \delta = c_2 \delta = \dots = c_n \delta \\ \implies c_1 &= c_2 = \dots = c_n = c \text{ so } P([x_{i-1}, x_i]) = c \delta \text{ for } i = 1, 2, \dots, n. \end{aligned}$$

Now, using property of probability, we have

$$\begin{aligned} P([x_0, x_1]) + P([x_1, x_2]) + \dots + P([x_{n-1}, x_n]) &= P(\Omega) = 1 \implies nc\delta = 1 \\ \implies c &= \frac{1}{n\delta} = \frac{1}{b-a} \implies P([x_{i-1}, x_i]) = \frac{\delta}{b-a} \text{ for } i = 1, 2, \dots, n. \end{aligned}$$

So, an equally likely or uniform law assumption for a continuous sample space is a valid one and produces a probability equal to the ratio of the length of the interval and the length of the sample space as follows:

$$P([a, b]) = \frac{\text{length of } [a, b]}{\text{length of } \Omega} = \frac{b-a}{\text{length of } \Omega}.$$

Since the probability of a point event occurring is zero, the probability of any interval

$$P([a, b]) = P((a, b]) = P([a, b)) = P((a, b)) = \frac{b - a}{\text{length of } \Omega}.$$

**Remark 2.5.** In continuous probabilistic modelling, we define probability of a subset of  $[0, 1]$  as the length of the subset. It is easy to compute of a subset in an interval or union of intervals. In real analysis, the concept of length of a set is generalized to a measure of the set and it is always not possible to come up with an explicit probability measure of every subset of a continuum sample space. That is, we need to introduce some kind of finiteness pattern in the continuum set, so that, a concrete probability measure can be defined similar to computation of probabilities in the remark (2.4). For the sake of simplicity, the finite pattern approach may lead to a well defined result that every measurable set in a continuum sample space is having a probability measure.

**Example 2.27** (Linear dartboard). Suppose one throws a dart at a linear dartboard as shown in Figure (2.8) and measures the horizontal distance from the bullseye or center at  $x = 0$ . We will then have a sample space  $\Omega = \{x \mid -\frac{1}{2} \leq x \leq \frac{1}{2}\}$  which is continuum in nature i.e. not countable. A possible approach is to assign probabilities to intervals (i.e. events in  $\Omega$ ) as opposed to sample points. If the dart is equally likely to land anywhere, then we could assign the interval  $[a, b]$  a probability equal to the length of the interval or

$$P([a, b]) = \frac{b - a}{\frac{1}{2} + \frac{1}{2}} = b - a \text{ for } -\frac{1}{2} \leq a \leq b \leq \frac{1}{2}.$$

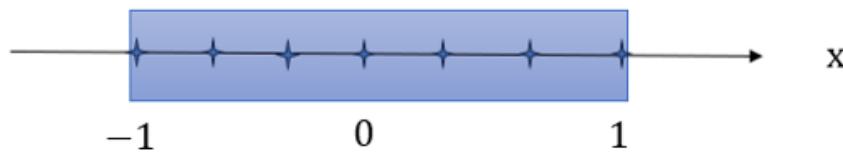


Figure 2.8: A linear dartboard.

**Example 2.28.** Question: a person always arrives at his job between 8:00 AM and 8:20 AM. He is equally likely to arrive any time within that period. What is the probability that he will arrive at 8:10 AM? What is the probability that he will arrive between 8:05 and 8:10 AM? Solution: since arrival time between 08:00 AM to 08:20 AM. And the person is arriving in equally likely manner in the interval. So, we have

$$(a.) P(\text{he will arrive at 8:10 AM}) = 0.$$

And

$$\begin{aligned} (b.) P(\text{he will arrive between 8:05 AM and 08:10 AM}) &= \frac{08 : 10 - 08 : 05}{08 : 20 - 08 : 00} \\ &= \frac{5}{20} = \frac{1}{4}. \end{aligned}$$

**Example 2.29 (Arrival time).** A person always arrives at his job between 8:00 AM and 8:20 AM. He is equally likely to arrive any time within that period. What is the probability that he will arrive at 8:10 AM? What is the probability that he will arrive between 8:05 and 8:10 AM?

Solution: since arrival time between 08:00 AM to 08:20 AM. And the person is arriving in equally likely manner in the interval. So, we have

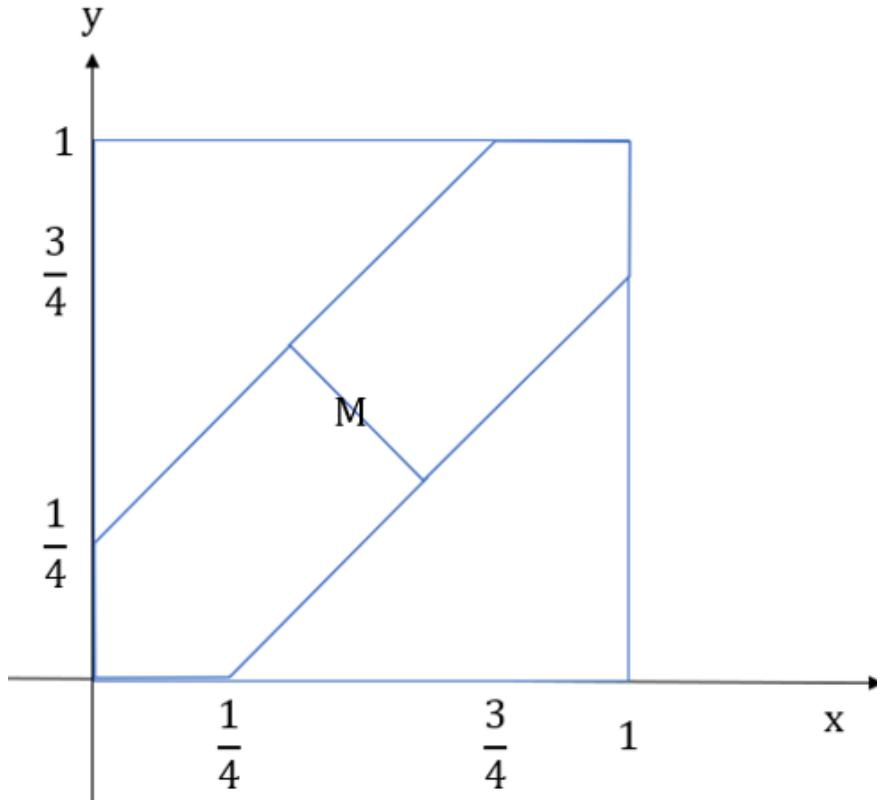
$$(a.) P(\text{he will arrive at 8:10 AM}) = 0.$$

And

$$\begin{aligned} (b.) P(\text{he will arrive between 8:05 AM and 08:10 AM}) &= \frac{08 : 10 - 08 : 05}{08 : 20 - 08 : 00} \\ &= \frac{5}{20} = \frac{1}{4}. \end{aligned}$$

**Example 2.30 (Romeo and Juliet meeting [Bertsekas and Tsitsiklis, 2008]).** Romeo and Juliet have a date on given time and each will arrive at the meeting with a delay between 0 and 1 hour with all the pair of delays are equally likely. The first to arrive wait 15 minutes and will leave the place if the other has not arrived. Compute the probability of meeting.

The sample space of the experiment is as follows:  $\Omega = \{(x, y) | 0 \leq x \leq 1, 0 \leq y \leq 1\}$ . Event of interest of the problem is  $A = \{(x, y) | |x - y| \leq \frac{1}{4}, 0 \leq x \leq 1, 0 \leq y \leq 1\}$  and is shaded in the figure (2.9).



**Figure 2.9:** The event  $A$  that Romeo and Juliet will arrive within 15 minutes of each other.

Thus, the desired probability is given by

$$P(A) = 1 - \frac{3}{4} \cdot \frac{3}{4} = 1 - \frac{9}{16} = \frac{7}{16}.$$

## 2.4 Conditional probability

In many real-world experiments, an outcome may not be completely random since we have some prior knowledge. For instance, (a.) knowing that it has rained the previous 2 days might influence our assignment of the probability of sunshine for the following day, (b.) knowing that from a population, a person's height exceeds 6 ft might influence assignment of probability that he weighs more than 100kg. It is the interaction between the original probabilities and the probabilities in light of prior knowledge that we wish to describe and quantify, leading to the concept of a conditional probability. So, it provides us with a way to reason about the outcome of an experiment, based on partial information.

**Definition 2.5** (Conditional probability). Suppose that we know outcome of an event  $A$  is within some given event  $B$ , then conditional probability of  $A$  given  $B$  with  $P(B) > 0$  is defined as below:

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

सप्रतिबंध प्रायिकता: सप्रतिबंध प्रायिकता ऐसी प्रायिकता होती है जिसमें हम एक घटना के होने के बाद दूसरी घटना के होने की प्रायिकता निकालते हैं।

प्रतिबंध: पु० [सं० प्रतिबन्ध् (बंधन)+घञ्] वह बंधन या सूचना जो किसी बात को कोई प्रायिक, स्वाभाविक या अधिकृत आचरण आदि करने से पहले ही लगाई गई हो।

**Restriction:** The restriction or information which is imposed on (something) before doing any usual, instinctive or authorized observance etc.

**Figure 2.10:** A self-explanatory meaning of conditional probability in Hindi with corresponding translation in English.

**Example 2.31.** Question: In three successive tosses of a fair coin, compute the conditional probability  $P(A|B)$  when  $A$  and  $B$  are the events

$$A = \{\text{more heads than tails come up}\}, B = \{\text{1st toss is a head}\}.$$

Answer: We have the sample space as a collection of all possible outcomes

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

and the events:

$$A = \{\text{more heads than tails come up}\}, B = \{\text{1st toss is a head}\}.$$

Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3}{4}.$$

**Example 2.32** ([Bertsekas and Tsitsiklis, 2008]). Question: a conservative design team, call it C, and an innovative design team, call it N, are asked to separately design a new product within a month. From past experience we know that:

- (a) The probability that team C is successful is 2/3.
- (b) The probability that team N is successful is 1/2.
- (c) The probability that at least one team is successful is 3/4.

If both teams are successful, the design of team N is adopted. Assuming that exactly one successful design is produced, then compute the probability that it was designed by team N.

Answer: There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams as follows:

$$\Omega = \{SS, SF, FS, FF\} \text{ where } SF \equiv C \text{ succeeds and } N \text{ fails.}$$

02: Events of interests and corresponding probabilities.

- a.  $A_1 \rightarrow C$  is successful i.e.  $A_1 = \{SS, SF\}$ , so  $P(A_1) = 2/3$ .
- b.  $A_1 \rightarrow N$  is successful i.e.  $A_2 = \{FS, FF\}$ , so  $P(A_2) = 1/2$ .
- c.  $A_3 \rightarrow$  at least one team is successful i.e.  $A_3 = \{SS, SF, FS\}$ , so  $P(A_3) = 3/4$ .

- d. From normalizing property of  $\Omega$ , we have  $P(\{SS, SF, FS, FF\}) = P(\Omega) = 1$ .

So, we have

$$P(SS) = 5/12, P(SF) = 1/4, P(FS) = 1/12, P(FF) = 1/4.$$

The the desired probability is given by

$$P(A) = P(\{SF\} | \{SF, FS\}) = \frac{P(SF)}{P(SF) + P(FS)} = \frac{1}{4}.$$

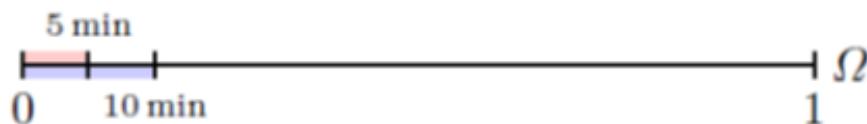
**Example 2.33** (bus arrival time). Question: this morning, I caught the bus whose arrival time is based on an uniform assumption (i.e. it is equally likely to come at any time) that bus always comes within at most one hour. Compute the probability that whether bus came within the first 5 minutes in the following two situation (a.) not any further information, (b.) the bus actually came in the first 10 minutes.

Solution: In the absence of any further information, you would say this is quite unlikely as

$$P\left(t \in \left[0, \frac{1}{12}\right]\right) = \frac{1}{12} \approx 0.083.$$

I give you a hint: I tell you that the bus actually came in the first 10 minutes. given this additional information, it seems much more likely that the bus came in the first 5 minutes than without this information in figure (2.11).

$$P\left(t \in \left[0, \frac{1}{12}\right] \mid \left[0, \frac{1}{6}\right]\right) = \frac{1}{2} = 0.5.$$



**Figure 2.11:** Bus arrival time conditioned on the bus actually came in the first 10 minutes.

### 2.4.1 Joint probability

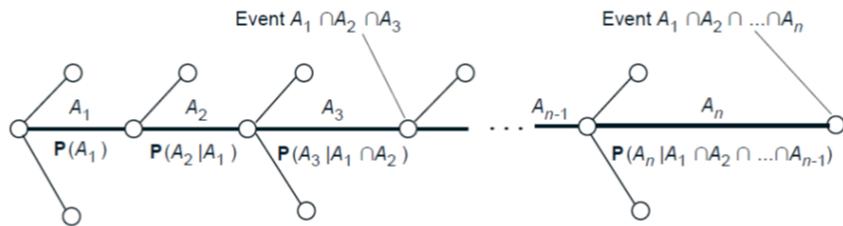
**Definition 2.6** (Joint probability [Bertsekas and Tsitsiklis, 2008]). In order to compute joint probability of two events A and B is a restatement of the definition of conditional probability (2.5) leads to a multiplication rule as:

$$P(A \cap B) = \begin{cases} P(A)P(B|A) & \text{if A occurs first} \\ P(B)P(A|B) & \text{if B occurs first.} \end{cases}$$

Furthermore, an event A which occurs if and only if each one of several events  $A_1, \dots, A_n$  has occurred, i.e.

$$A = A_1 \cap \dots \cap A_n \rightarrow P(A) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1, \dots, A_{n-1}).$$

The occurrence of A can be viewed as an occurrence of  $A_1$ , followed by the occurrence of  $A_2$ , then of  $A_3$ , etc, and it is visualized as a path on the tree with n branches, corresponding to the events  $A_1, \dots, A_n$  in (2.12).



**Figure 2.12:** Visualizing occurrence of A through joint occurrence of  $A_1, \dots, A_n$  sequentially.

**Example 2.34.** Question: In a factory there are 100 units of a certain product, 5 of which are defective. When we pick three units at random, compute the probability that none of them are defective.

Solution: let  $A_i$  is the event that the ith chosen unit is not defective, for  $i = 1, 2, 3$ . Then, we need to compute  $P(A_1 \cap A_2 \cap A_3)$  as the desired probability of interest.

Now, we have

$$P(A_1) = \frac{95}{100} \text{ from uniform assumption.}$$

Given that the first chosen item was good, the second item will be chosen from 94 good units and 5 defective units, thus

$$P(A_2|A_1) = \frac{94}{99} \text{ from uniform assumption conditioned on } A_1.$$

Given that the first and second chosen items were okay, the third item will be chosen from 93 good units and 5 defective units, thus

$$P(A_3|A_1, A_2) = \frac{93}{98} \text{ from uniform assumption conditioned on } A_1, A_2.$$

So, the desired probability is given by

$$\begin{aligned} P(A) &= P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \\ &= \frac{95}{100} \cdot \frac{94}{99} \cdot \frac{93}{98} = 0.8560. \end{aligned}$$

**Definition 2.7 (Independent events).** Consider two events A and B, if chance of occurrence of one first does not effect the chance of occurrence of the other i.e. conditioned on the outcome of one event does not change the probability of the other event, then these two events are independent to each other. Mathematically,

$$P(B|A) = P(B) \text{ or } P(A|B) = P(A) \implies P(A \cap B) = P(A)P(B).$$

It can generalized to a sequence of events namely  $A_1, A_2, \dots, A_n$ : the events  $A_1, \dots, A_n$  are independent if

$$P(A_i|A_{j_1}, \dots, A_{j_k}) = P(A_i) \text{ for } i \neq j_1 \neq \dots \neq j_k, 1 \leq k \leq n.$$

Furthermore, conditioned on an event C, the events A and B are called conditionally independent if

$$\begin{aligned} P(B|A, C) &= P(B|C) \text{ or } P(A|B, C) = P(A|C) \implies \\ P(A \cap B|C) &= P(A|C)P(B|C). \end{aligned}$$

**Example 2.35.** Backdrop: Consider two independent fair coin tosses, in which all of our possible outcomes are equally likely. Let

$$A_1 = \{1st \text{ toss is a head.}\}, A_2 = \{2nd \text{ toss is head.}\}, \\ B = \{\text{the two tosses have different results.}\}$$

Question: show that (a.)  $A_1$  and  $A_2$  are independent, (b.) but conditioned on  $B$ ,  $A_1$  and  $A_2$  are dependent.

Answer: we have

$$A_1 = \{HT, HH\}, A_2 = \{TH, HH\}, B = \{HT, TH\} \implies P(A_1) = \frac{1}{2}, \\ P(A_2) = \frac{1}{2}, P(B) = \frac{1}{2}, P(A_1 \cap A_2) = \frac{1}{4} \\ P(A_1|B) = \frac{1}{2}, P(A_2|B) = \frac{1}{2}, P(A_1 \cap A_2|B) = 0.$$

Then

$$P(A_1 \cap A_2) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1)P(A_2) \implies \\ A_1 \text{ and } A_2 \text{ are independent.}$$

Again

$$P(A_1 \cap A_2|B) = 0 \neq P(A_1|B)P(A_2|B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \implies \\ A_1|B \text{ and } A_2|B \text{ are dependent.}$$

### 2.4.2 Total probability

**Definition 2.8** (Mutually disjoint events). Consider any two sets  $A$  and  $B$ . Then  $A$  and  $B$  are mutually disjoint if

$$A \cap B = \varnothing.$$

Moreover, a list of events namely  $E_1, E_2, E_3, \dots$  are mutually disjoint if the intersection of any pair of events is empty, that is,

$$E_i \cap E_j = \varnothing \text{ for } i \neq j, i, j = 1, 2, 3, \dots$$

**Definition 2.9** (Mutually exclusive events). Consider two events  $A$  and  $B$  in a probabilistic modelling of a random experiment. Then  $A$  and  $B$

are mutually exclusive if

$$P(A \cap B) = 0 \text{ i.e. joint occurrence of A and B is impossible.}$$

That is, happening of one in a single trial, kills the happening of the other (or others) in the same trial. For example, a coin cannot land with head and tail (or a die cannot land with several faces up), so they are exclusive outcomes.

Moreover, a list of events namely  $E_1, E_2, E_3, \dots$  are mutually exclusive if the joint occurrence of any pair of events is impossible, that is,

$$P(E_i \cap E_j) = 0 \text{ for } i \neq j, i, j = 1, 2, 3, \dots$$

And, for mutually exclusive events  $E_1, E_2, E_3, \dots$ , we have

$$P(E_1) + P(E_2) + P(E_3) + \dots \leq 1 = P(\Omega).$$

**Definition 2.10** (collectively exhaustive). A list of events  $E_1, E_2, E_3, \dots$  in a probabilistic modelling of a random experiment is collectively exhaustive if

$$E_1 \cup E_2 \cup E_3 \cup \dots = \Omega \implies P(E_1 \cup E_2 \cup E_3 \cup \dots) = 1 = P(\Omega).$$

Moreover, if a list of events  $E_1, E_2, E_3, \dots$  are mutually exclusive and collectively exhaustive, then

$$P(E_1) + P(E_2) + P(E_3) + \dots = 1 = P(\Omega).$$

**Example 2.36.** Consider a random experiment with the sample space:  $\Omega = \{1, 2, 3, 4\}$ , and define a few events as:

$$A_1 = \{1, 2\}, A_2 = \{3\}, B_1 = \{1, 2, 3\}, B_2 = \{3, 4\}.$$

Then

- a.  $A_1 = \{1, 2\}$  and  $A_2 = \{3\}$  are mutually exclusive but not collectively exhaustive.
- b.  $B_1 = \{1, 2, 3\}$  and  $B_2 = \{3, 4\}$  are collectively exhaustive but not mutually exclusive.

- c.  $\{A_1, B_2\} = \{1, 2, 3, 4\}$  is a collection of mutually exclusive and collectively exhaustive events, and is thus it forms a partition of the sample space  $\Omega$ .

**Theorem 2.3** (Mutually disjoint vs mutually exclusive). In a probabilistic modelling of a random experiments, mutually disjoint events are also mutually exclusive, but converse is not true.

*Proof.* Consider two mutually disjoint events A and B in a probabilistic modelling of a random experiment. Then

$$A \cap B = \varphi \implies P(A \cap B) = P(\varphi) = 0.$$

Since  $P(A \cap B)$ , so the mutually disjoint events A and B are also mutually exclusive.

A counter example of the converse: consider points in the square with each coordinate uniformly distributed from 0 to 1. Let A be the event where the x-coordinate is 0, and B be the event that the y-coordinate is 0. Then

$$A \cap B = \{(0, 0)\} \neq \varphi, \text{ but } P(A \cap B) = 0.$$

So A and B are not mutually disjoint, but they are mutually exclusive. Another counter example: consider a sample space be  $\Omega = \{x, y, z\}$  with probabilities  $P(\{x\}) = 0, P(\{y\}) = 1/2$ , and  $P(\{z\}) = 1/2$ . If  $A = \{x, y\}$  and  $B = \{x, z\}$ , then  $A \cap B = \{x\}$ , but  $P(A \cap B) = P(\{x\}) = 0$ . That is, they are mutually exclusive but not disjoint.  $\square$

**Definition 2.11** (Total probability). Consider a sample space  $\Omega$  with a partition  $\{B_1, \dots, B_n\}$  i.e.

- a. mutually exclusive:

$$B_i \cap B_j = \varphi \text{ for any pair of } B_i, B_j, i \neq j.$$

- b. collectively exhaustive:

$$B_1 \cup B_2 \cup \dots \cup B_n = \Omega.$$

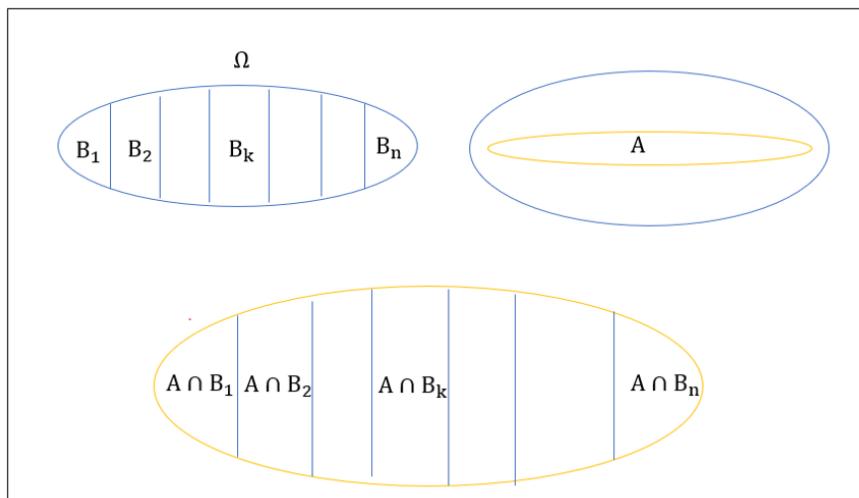
Suppose  $A \subset \Omega$  be an event in the experiment, then  $\{B_1 \cap A, \dots, B_n \cap A\}$  introduces a partition of A as (in figure (2.13)):

a. mutually exclusive:

$$(B_i \cap A) \cap (B_j \cap A) = \varphi \text{ for any pair of } B_i, B_j, i \neq j.$$

b. collectively exhaustive:

$$B_1 \cap A \cup B_2 \cap A \cup \dots \cup B_n \cap A = A.$$



**Figure 2.13:** Visualizing a partition of event A through partition of the sample space  $\Omega$ .

Then,

$$\begin{aligned} P(A) &= P(A \cap B_1 \cup \dots \cup A \cap B_n) = \sum_{i=1}^n P(A \cap B_i) \\ &= \sum_{i=1}^n P(B_i)P(A|B_i). \end{aligned}$$

**Example 2.37.** Question: Three machines make parts at a factory. Suppose, we know the following about the manufacturing process that Machine 1 ( $M_1$ ) makes 60% of the parts, Machine 2 ( $M_2$ ) makes 30% of

the parts, Machine 3 ( $M_3$ ) makes 10% of the parts. Of the parts  $M_1$  makes 7% are defective, of the parts  $M_2$  makes 15% are defective, of the parts  $M_3$  makes 30% are defective. Compute the probability of defective products produced by the factory.

Answer: the desired probability of defective product produced by the factory is given by

$$\begin{aligned} P(D) &= P(M_1)P(D|M_1) + P(M_2)P(D|M_2) + P(M_3)P(D|M_3) \\ &= 0.6 \cdot 0.07 + 0.3 \cdot 0.15 + 0.1 \cdot 0.3 = 0.117. \end{aligned}$$

### 2.4.3 Bayes rule

**Definition 2.12** (Bayes rule). Consider two events A and B with  $P(A) > 0, P(B) > 0$ , then Bayes rule is a restatement of the definition of conditional probability (2.5) together with total probability (2.11) as:

$$\begin{aligned} P(A \cap B) &= \begin{cases} P(A)P(B|A) & \text{if A occurs first} \\ P(B)P(A|B) & \text{if B occurs first.} \end{cases} \\ \implies P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{\sum_{i=1}^n P(B_i)P(A|B_i)} \end{aligned}$$

where, events  $\{B_1, \dots, B_n\}$  together is a partition of the sample space  $\Omega$ , and hence it brings a partition of the event A as  $\{A \cap B_1, \dots, A \cap B_n\}$ . Furthermore, the probabilities of effect causes  $\{B_1, \dots, B_n\}$  conditioned on effect A, can be reverse calculated as

$$P(B_k|A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(B_k)P(A|B_k)}{\sum_{k=1}^n P(B_i)P(A|B_i)}.$$

**Example 2.38** (01: Medical diagnostics for a rare disease i.e. a false positive). Let A be the event that the test comes back positive, and B be the event that you have the disease. Our modelling assumptions are: 1. 95% of patients with the disease test positive i.e.  $P(A|B) = 0.95$ , 2. 2% of patients without the disease test positive i.e.  $P(A|B^c) = 0.02$ , 3. one in a thousand people have the disease i.e.  $P(B) = 0.001$ . Compute the probability  $P(B|A) \rightarrow$  you have the disease given that the test comes back positive.

Answer: we have the following probabilistic data:

$$P(A|B) = 0.95, P(A|B^c) = 0.02, P(B) = 0.001.$$

Now, in order to compute the desired probability  $P(B|A)$ , we apply Bayes rule as :

$$\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} = \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.02} \\ &\approx 0.045. \end{aligned}$$

**Example 2.39** (Airport security). Question: on an airport all passengers are checked carefully. Let  $T$  be the random variable indicating whether somebody is a terrorist (i.e.  $t = 1$ ) or not (i.e.  $t = 0$ ) and  $A$  be the variable indicating arrest (i.e.  $A = a \in \{0, 1\}$ ). A terrorist shall be arrested with probability  $P(A = 1|T = 1) = 0.98$ , a non-terrorist with probability  $P(A = 1|T = 0) = 0.001$ . One in hundred thousand passengers is a terrorist, (i.e.  $P(T = 1) = 0.00001$ ). Compute the probability that an arrested person actually is a terrorist.

Answer: using Bayes theorem, we have

$$\begin{aligned} P(T = 1|A = 1) &= \frac{P(T = 1)P(A = 1|T = 1)}{P(A = 1)} \\ &= \frac{P(T = 1)P(A = 1|T = 1)}{P(T = 0)P(A = 1|T = 0) + P(T = 1)P(A = 1|T = 1)} \\ &= \frac{0.00001 \cdot 0.98}{(1 - 0.00001) \cdot 0.001 + 0.00001 \cdot 0.98} \approx 0.01 \end{aligned}$$

**Remark 2.6** (Screening test for a disease). One thousand in a 100 thousand people have the disease i.e. disease rate is 1%. 99% accurate doesn't really give us information about the disease. So, we use the following terms:

- a. Sensitivity - the odds that the test will be positive if you have the disease. A 98% sensitive means, it will correctly identify 980 of 1000 people with the disease i.e. it also incorrectly claim 20 = 1000 – 980 people with the disease don't have it,
- b. Specificity - the odds that the test will be negative if you lack the disease. A 99% specific means, it correctly identify 98,010

of 99,000 without the disease i.e. it also incorrectly claim  $990 = 99000 - 98010$  people without the disease have it.

- c. Positive predictive value - the odds that the test will correctly predict you have the disease, if you test positive. So, out of  $1970 = 980 + 990$  people (true positive and false positive) who test positive, 980 have the disease. Thus, our positive predictive value is  $1000/1970 = 50.76\%$ .
- d. Negative predictive value - the odds that the test will correctly predict you lack the disease, if you test negative. So, Out of  $98030 = 98010 + 20$  who test negative (false negative and true negative), 99000 do not have the disease. Thus, our negative predictive value is  $98030/99000 = 99.02\%$ .

In this case, this test is first-rate for determining who lacks the disease. The 1970 (true positive and false positive), who test positive can be tested to confirm they do have the disease, whereas those who tested negative (false negative and true negative) need no further tests.

**Remark 2.7.** The accuracy of a DNA test would be in its ability to predict a match vs non-match between two samples of blood. For example, I draw blood from one person into two tubes, then perform the test and see if it matches. Repeat a bunch and divide the positives by the total, that's the sensitivity. Then draw blood from two random donors, repeat, and divide the number of negatives over the total. That's the specificity. They are testing whether the machine can tell  $A = A$  and  $\neq B$ , not whether a random sample of all the humans on the planet will match your blood specifically.

**Example 2.40 (Drug test).** Question: a drug test (say T) has 1% false positives (i.e. 1% of those not taking drugs show positive in the test), and 5% false negatives (i.e. 5% of those taking drugs test negative). Suppose that 2% of those tested are taking drugs. Compute the probability that somebody who tests positive is actually taking drugs (say D).

Answer: we define here

- a.  $T = p \rightarrow$  Test positive,

- b.  $T = n \rightarrow$  Test negative,
- c.  $D = p \rightarrow$  person takes drug,
- d.  $D = n \rightarrow$  person does not take drugs

Now, we have

- i. false positive i.e. test positive given that person does not take drug:  $P(T = p|D = n) = 0.01$
- ii. false negatives i.e. test negative given that person takes drug:  $P(T = n|D = p) = 0.05 \implies P(T = p|D = p) = 0.95$  : true positives i.e. test positive given that person takes drug.
- iii. drug consuming rate:  $P(D = p) = 0.02 \implies P(D = n) = 0.98$ .

Then, using Bayes theorem, the probability that somebody who tests positive is actually taking drugs computed as:

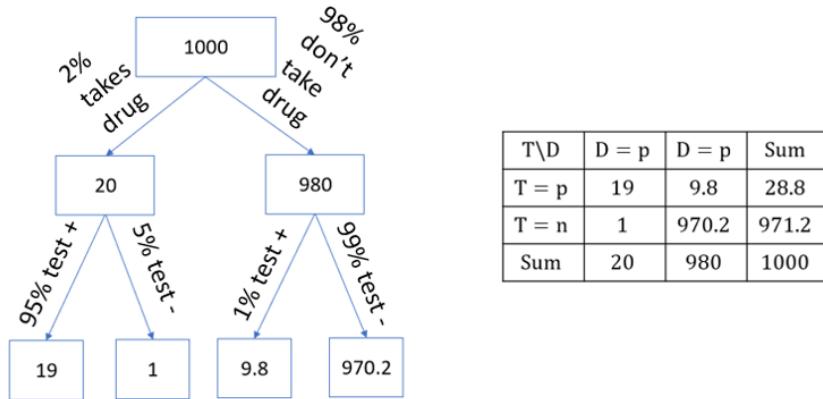
$$\begin{aligned} P(D = p|T = p) &= \frac{P(D = p)P(T = p|D = p)}{P(T = p)} \\ &= \frac{P(D = p)P(T = p|D = p)}{P(D = n)P(T = p|D = n) + P(D = p)P(T = p|D = p)} \\ &= \frac{0.02 \cdot 0.95}{0.98 \cdot 0.01 + 0.02 \cdot 0.95} \approx 0.66. \end{aligned}$$

An alternative way to solve this exercise is using decision trees. Let's assume there are 1000 people tested. Then we have following decision tree:

So, from the contingency table (2.14), we have

$$\frac{\text{taking drug and tests positive}}{\text{all positive tests}} = \frac{19}{28.8} \approx 0.66.$$

There is a chance of only two thirds that someone with a positive test is actually taking drugs.



**Figure 2.14:** In left plot, there is a decision tree of 100 tested people and in the right plot there is a contingency table for tested people.

**Example 2.41 (Oral exam).** In an oral exam you have to solve exactly one problem, which might be one of three types, A, B, or C, which will come up with probabilities 30%, 20%, and 50%, respectively. During your preparation you have solved 9 of 10 problems of type A, 2 of 10 problems of type B, and 6 of 10 problems of type C. Then, (a) compute the probability that you will solve the problem of the exam, (b.) given you have solved the problem, compute the probability that it was of type A.

Answer: the probability to solve the problem of the exam is the probability of getting a problem of a certain type times the probability of solving such a problem, summed over all types. This is known as the total probability.

$$\begin{aligned}
 P(\text{solved}) &= P(\text{solved}|A)P(A) + P(\text{solved}|B)P(B) + P(\text{solved}|C)P(C) \\
 &= \frac{9}{10} \cdot 30\% + \frac{2}{10} \cdot 20\% + \frac{6}{10} \cdot 50\% = \frac{27}{100} + \frac{4}{100} + \frac{30}{100} = 0.61.
 \end{aligned}$$

Furthermore, using Bayes theorem, we have

$$P(A|\text{solved}) = \frac{P(A)P(\text{solved}|A)}{P(\text{solved})} = \frac{30\% \cdot \frac{9}{10}}{0.61} = 0.442.$$

# 3

---

## Probability distribution

---

Till now, we have seen in section (2.2) that in principle of probabilistic modelling of a random experiment in terms of outcomes and event and carried out through the following three basic concepts:

- a. Sample space → as a collection of all possible outcomes of a random experiment.
- b. Event  $\omega$  as a subset of the sample (i.e. a collection of some possible outcomes) defined by an statement.
- c. Probability measure → as a map  $P : \Sigma \rightarrow [0, 1]$  such that a.  $P(\emptyset) = 0, P(\Omega) = 1$ , b.  $0 \leq P(A) \leq 1$  for any event  $A \in \Sigma$ , c.  $P(A \cup B) = P(A) + P(B)$  provided A and B are mutually disjoint (i.e.  $A \cap B = \emptyset$ ). In summary, compute the probability of an event by breaking it into disjoint pieces whose probabilities are summed.

But in practice, everything we measure in real life is random in the sense (i.e. if we perform the experiment again we will not get the same measured value). So, every measurement is a random variable ( $X : \Omega \rightarrow \Omega_X \subset \mathbb{R}, X^{-1}(B) \in \Sigma_\Omega$  for every  $B \in \mathcal{B}$ ) that provides a

concrete/computational representation  $\Omega_X$  of the sample space  $\Omega$  in a numeric/quantified platform.

We have already observed that an interesting and generic mathematical principle: to study, analyse and understand a space (with some structure) is extremely advantageous via study of functions from the space to a target space such that functions respect the structure of the space. For example:

1. An  $m \times n$  matrix as a linear transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ .
2. A linear representation of a group to a general linear group of matrices.
3. A rational function from a variety to an algebraic variety.
4. A random variable  $X : \Omega \rightarrow \Omega_X \subset \mathbb{R}$ ,  $X^{-1}(B) \in \Sigma_\Omega$  for every  $B \in \mathcal{B}$ .

This observation makes random variable much important for the computation of randomness in a real word situation.

### 3.1 Random variable

In Bernoulli trials, we are interested only in number of successes rather than successes or failure. In practice, everything in applied/computational probability is about random variables and their probability distributions.

**Definition 3.1 (Informal).** A random variable translates an outcome of nature in a given random experiment to some random number ( i.e. a number with a probability of occurrence). For example, distance is a function of a pair of points, the perimeter of a triangle is a function of triangles. In summary, a random variable translates non-numerical sample space  $\Omega$  of a random experiment to numerical sample space  $\Omega_X$ .

**Definition 3.2 (Formal).** A random variable  $X$  is a  $(\mathbb{R}, \mathcal{B})$ – valued map  $X : \Omega \rightarrow \mathbb{R}_+$  such that  $X^{-1}(B) = \{\omega | X(\omega) \in B\} \subset \Sigma$  for every  $B \in \mathcal{B}$  (i.e.  $X^{-1}(B)$  is a measurable set w.r.t probability measure P), where  $\Omega$  is a sample space of a given random experiment and  $\mathbb{R}$  endures a

Borel  $\sigma$ -algebra  $\mathcal{B}$  on it generated by open intervals. Alternatively, a random variable is a measurable function  $X : (\Omega, \Sigma, P) \rightarrow (\mathbb{R}, \mathcal{B})$ .

**Definition 3.3** (Probability distribution). Probability distribution of a random variable (3.2)  $X$  is defined as a probability measure  $p_X = P \circ X^{-1} : (\mathbb{R}, \mathcal{B}) \rightarrow [0, 1]$  such that

$$P \circ X^{-1}(A) = P(X^{-1}(A)) = P(X \in A) = P(\{\omega | X(\omega) \in A\}).$$

It describes the likelihood of occurrence of the possible value of a random variable.

प्रायिकता वितरण : एक संखियकीय फलन जो किसी दिए गए सीमा के भीतर यादृच्छिक चर के लिए सभी संभावित मूल्यों और प्रायिकता का वर्णन करता है।

प्रायिक : वि० [सं० प्राय+ठक्-इक] [भाव० प्रायिकता] जो नियमित रूप से या सदा तो नहीं फिर भी बीच-बीत में प्रायः होता रहता हो।

Which does not happen regularly or always, but still happens often from time to time.

वितरण: पुं० [सं० विर्तु (पार करना)+ल्युट-अन] उत्पत्ति के फलस्वरूप होनेवाली प्राप्ति का उत्पत्ति के साधनों में बाँटना।

Distribution: Distribution of resources resulting from origination among the means of origination.

**Figure 3.1:** A self-explanatory meaning of probability distribution in Hindi with corresponding translation in English.

**Remark 3.1.** Once we get to more complicated examples, it can quickly become impractical to specify explicitly. Although the concept of a probability space  $(\Omega, \Sigma, P)$  underlies everything, in practice it will be rare that we think about itself, instead we will talk directly about events and their probabilities, and random variables and their distributions (and we can do that without assuming any particular structure for  $\Omega$ ).

## 3.2 Discrete distribution

**Definition 3.4** (Discrete random variable). A measurable function  $X : \Omega = \{\omega_1, \omega_2, \dots, \} \rightarrow \Omega_X = \{x_1, x_2, \dots\} \subset \mathbb{R}$  with  $X^{-1}(B) \in \Sigma_\Omega =$

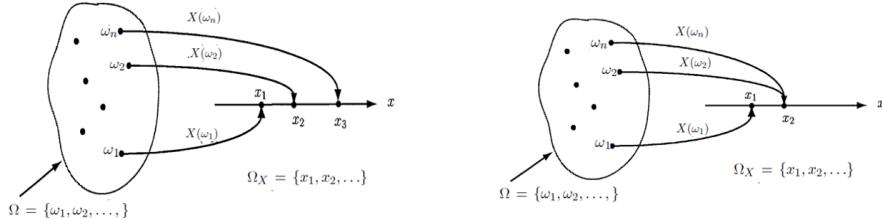
$\{X(\omega_1), X(\omega_2), \dots, X(\omega_n)\} \forall B \in \mathcal{B}$  such that

- a.  $X^{-1}(x) = \{\omega | X(\omega) = x\} \in \Sigma_\Omega$ ,
- b.  $\Omega_X = \{x_k\}_{k \in \mathbb{N}}$  i.e.  $\Omega_X$  is at most countable.
- a'.  $\Omega_X = \{x_k\}_{k \in \mathbb{N}} \implies B = \{x\} \implies X^{-1}(B) = X^{-1}(x) = \{\omega | X(\omega) = x\} \equiv \{X = x\}$ .

is called a discrete random variable and its probability distribution is known as probability mass function. An event of a discrete random variable  $X$  is described as below:

$$X \in B = \{x\} \equiv \{X = x\} \equiv \{\omega | X(\omega) = x\}.$$

Here,  $\{X = x\} \rightarrow$  the event that the random variable  $X$  takes the given value  $x$ .



**Figure 3.2:** Discrete random variable as a mapping of at most countably sample space into a sequence of real numbers.

**Definition 3.5 (Probability mass function).** Probability distribution of a discrete random variable (DRV)  $X$  is known as a probability mass function (PMF) of the DRV and defined as follows:

$$p_X(x) = P(X = x) = P(\{\omega | X(\omega) = x\}).$$

The PMF  $p_X$  satisfies the following properties:

- a.  $p_X(x) \in [0, 1]$ ,
- b.  $\sum_x p_X(x) = 1$ ,
- c.  $P(X \in B) = \sum_{x \in B} p_X(x)$ .

### 3.2.1 Various discrete distribution

**Example 3.1** (Bernoulli distribution). A random variable  $X$  (with an event  $\{X = x\} \equiv \{\omega = H \text{ or } T | X(\omega) = x\}$ ) has a Bernoulli distribution with a parameter  $p \in [0, 1]$  if

$$p_X(x) = P(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases} = p^x(1 - p)^{1-x} \text{ for } x \in \{0, 1\}.$$

So,  $X$  is a Bernoulli random variable and we write  $X \sim Ber(p)$  (see in (3.3)). e.g. a.  $X = I_A \rightarrow$  an indicator function for an event  $A$ , b.  $X \rightarrow \# \text{ successes in a Bernoulli trial}$ .

The Bernoulli distribution is used to model, for example, the outcome of a Bernoulli trial with 1 representing successes and 0 representing failure. It is also a basic building block for other classical discrete distributions.

In the experiment of a single fair coin toss with a probability of getting head is  $p$ , if we are interested only in whether the outcome is head or tail, then we define a Bernoulli random variable as follows:

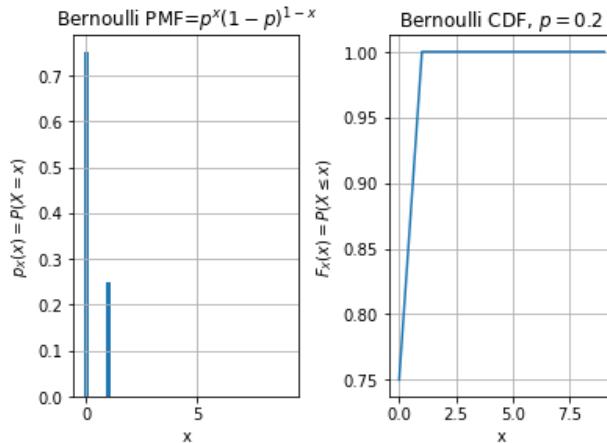
$$X(\omega) = \begin{cases} 0 & \text{if } \omega = T \\ 1 & \text{if } \omega = H. \end{cases}$$

With the following Bernoulli distribution as a PMF of the  $X$ :

$$\begin{aligned} p_X(0) &= p(X = 0) = \sum_{X^{-1}(0) \in \Sigma} P(\{\omega | X(\omega) = 0\}) = P(T) = 1 - p, \\ p_X(1) &= p(X = 1) = \sum_{X^{-1}(1) \in \Sigma} P(\{\omega | X(\omega) = 1\}) = P(H) = p. \end{aligned}$$

In the experiment of a single fair die toss if we are interested only in whether the outcome is even or odd, then we define a Bernoulli random variable as follows:

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = 1, 3, 5 \\ 1 & \text{if } \omega = 2, 4, 6. \end{cases}$$



**Figure 3.3:** Bernoulli probability mass function for  $p = 0.25$ .

With the following Bernoulli distribution as a PMF of the  $X$ :

$$p_X(0) = p(X = 0) = \sum_{X^{-1}(0) \in \Sigma} P(\{\omega | X(\omega) = 0\}) = 3/6 = 1/2,$$

$$p_X(1) = p(X = 1) = \sum_{X^{-1}(1) \in \Sigma} P(\{\omega | X(\omega) = 1\}) = 3/6 = 1/2.$$

Here, we can draw a remark that once the PMF has been specified all subsequent probability calculations can be based on it, without referring back to the original sample space  $S$ . Specifically, for an event  $A$  defined on  $S_x$  the probability is given by

$$P(X \in A) = \sum_{i:x_i \in A} p_X(x_i).$$

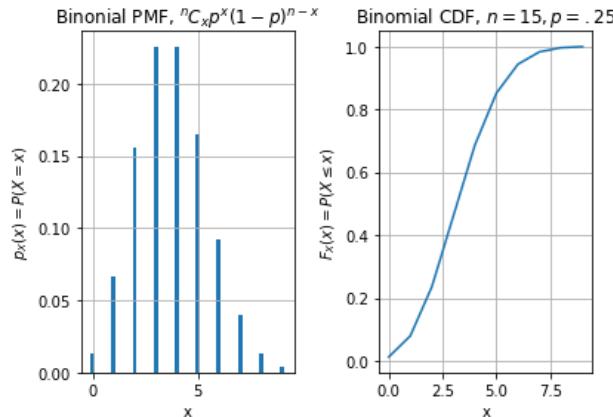
Consider a die whose sides have been labelled with two sides having 1 dot, two sides having 2 dots, and two sides having 3 dots. Hence,  $\Omega = \{\omega_1 = 1, \omega_2 = 1, \omega_3 = 2, \omega_4 = 2, \omega_5 = 3, \omega_6 = 3\}$ . Then if we are interested in the probabilities of the outcomes displaying either 1, 2, or 3 dots, we would define a random variable as

**Example 3.2** (Binomial distribution). A random variable  $X$  (with an event  $\{X = x\} \equiv \{\omega \in \{H, T\}^n | X(\omega) = x = \#H\}$ ) has a Binomial

distribution with a parameter  $p \in [0, 1]$  and  $n \in \mathbb{N}$  if

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x \in \{0, 1, 2, \dots, n\}.$$

So,  $X$  is a Binomial random variable and we write  $X \sim \text{Bin}(n, p)$  (see in (3.4)). e.g.  $X \rightarrow \#$  successes in  $n$  Bernoulli trials with  $p$  as a probability of a success. Probability of a particular sequence of  $n$  Bernoulli trials with  $x$  successes is  $p^x(1 - p)^{n-x}$  and there are exactly  $\binom{n}{x}$  such sequences.



**Figure 3.4:** Binomial probability mass function with its cumulative distribution for  $n = 15, p = 0.25$ .

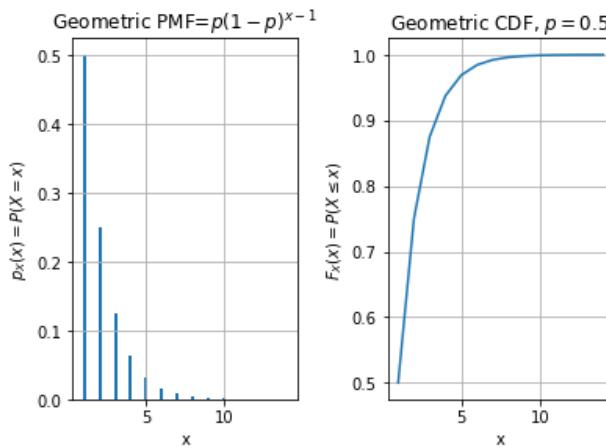
The location of the maximum of the PMF can be shown to be given by  $\arg \max_x \left( \binom{n}{x} p^x (1 - p)^{n-x} \right) = [(n+1)p]$  (see [Feller, 1968]), where  $[x]$  denotes the largest integer less than or equal to  $x$ .

**Example 3.3** (Geometric distribution). A random variable  $X$  (with an event  $\{X = x\} \equiv \{\omega \in \{H, T\}^\infty | X(\omega) = x = \# \text{ Bernoulli trials till first } H\}$ ) has a Geometric distribution with a

parameter  $p \in [0, 1]$  if

$$p_X(x) = P(X = x) = p(1 - p)^{x-1} \text{ for } x = 1, 2, \dots$$

So,  $X$  is a geometric random variable and we write  $X \sim Geo(p)$  (see in (3.5)). e.g.  $X \rightarrow \#$  Bernoulli trials till first success.

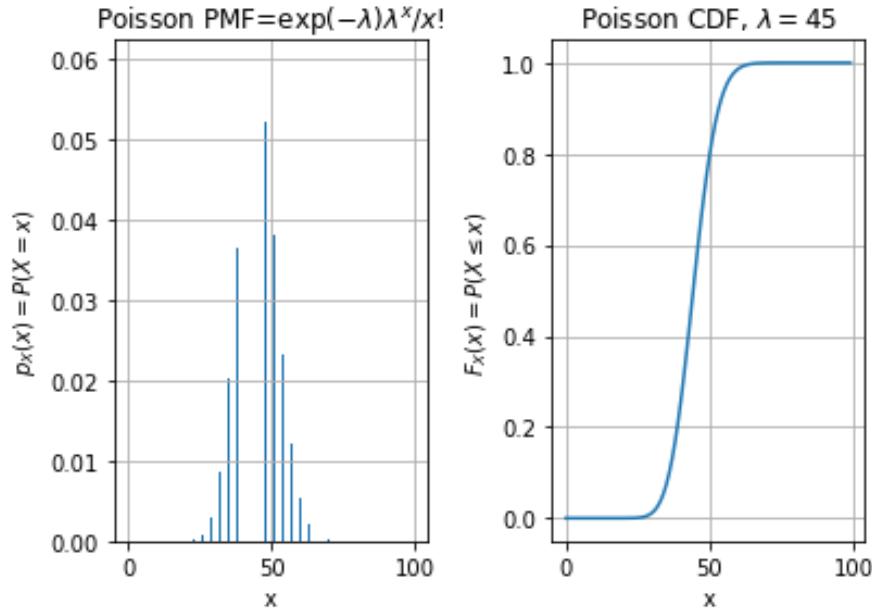


**Figure 3.5:** Geometric probability mass function for  $p = 0.25$ .

**Example 3.4** (Poisson distribution). A random variable  $X$  has a Poisson distribution with a parameter  $\lambda = np$ , where  $n$  is possibly large and  $p$  is possibly very small, if

$$\begin{aligned} p_X(x) = P(X = x) &= \lim_{n \rightarrow \infty} \left( \binom{n}{x} p^x (1-p)^{n-x} \right) \\ &= \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \dots \end{aligned}$$

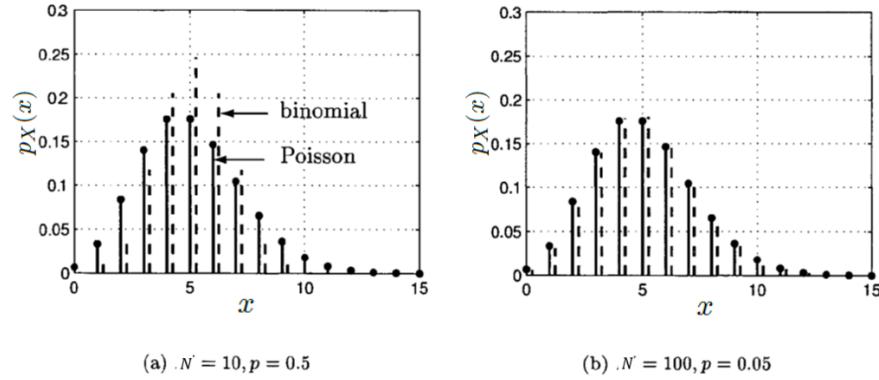
So,  $X$  is a Poisson random variable and we write  $X \sim Pois(\lambda)$  (see in (3.6)). e.g.  $X \rightarrow \#$  successes in Bernoulli trials per unit interval (or within a given interval of time or space).



**Figure 3.6:** The Poisson probability mass function with its cumulative distribution for  $\lambda = 75$ .

**Remark 3.2** (Poisson approximation of binomial distribution). If in a binomial PMF, we let  $n \rightarrow \infty$  as  $p \rightarrow 0$  such that the product  $\lambda = np$  remains constant, then  $Bin(n, p) \rightarrow Pois(\lambda)$ . Note that  $\lambda = np$  represents the expected or average number of successes in  $n$  Bernoulli trials, hence, by keeping the average number of successes fixed but assuming more and more trials with smaller and smaller probabilities of success on each trial, we are led to a Poisson PMF (see in (3.7)).

**Example 3.5** (Hyper-geometric distribution [Feller, 1968]). In a population of  $n$  elements  $n_1$  are red and  $n_2$  are black. A group of  $r$  elements is chosen at random. We seek the probability  $p_X(k)$  that the group so chosen will contain exactly  $k$  red elements. Here  $k$  can be any integer



**Figure 3.7:** The Poisson approximation to the binomial probability mass function.

between zero and  $n_1$  or  $r$ , whichever is smaller. We note that the chosen group contains  $k$  red and  $r - k$  black elements. The red ones can be chosen in  $\binom{n_1}{k}$  different ways and the black ones in  $\binom{n-n_1}{r-k}$  ways. Since any choice of  $k$  red elements may be combined with any choice of black ones, so we have

$$p_X(k) = P(X = k; n, n_1, r) = \frac{\binom{n_1}{k} \binom{n - n_1}{r - k}}{\binom{n}{r}}$$

The hyper-geometric distribution can easily be generalized to the case where the original population of size  $n$  contains several classes of elements. For example, let the population contain three classes of sizes  $n_1, n_2$ , and  $n - n_1 - n_2$  respectively. If a sample of size  $r$  is taken, the probability that it contains  $k_1$  elements of the first,  $k_2$  elements of the second, and  $r - k_1 - k_2$  elements of the last class is given by

$$p_X(k) = P(X = k; n, n_1, n_2, r) = \frac{\binom{n_1}{k_1} \binom{n_2}{k_2} \binom{n - n_1 - n_2}{r - k_1 - k_2}}{\binom{n}{r}}$$

In Bridge game: the population of 52 cards consists of four classes, each of thirteen elements. The probability that a hand of thirteen cards

consists of five spades, four hearts, three diamonds, and one club is

$$p_X(k) = P(X = (5, 4, 3, 1); 52, 13, 13) = \frac{\binom{13}{5} \binom{13}{4} \binom{13}{3} \binom{13}{1}}{\binom{52}{13}}$$

**Example 3.6** (Pascal or negative binomial distribution [Feller, 1968]). In a sequence of independent Bernoulli trials, let the random variable  $X$  be the trial at which the  $r$ th success occurs, where  $r$  is a fixed integer. Then  $X$  has a Pascal distribution as follows:

$$p_X(x) = \binom{x}{r-1} p(1-p)^{x-r}.$$

Consider a random variable  $Y \rightarrow$  number of failures before  $r$ th success. This formulation is statistically equivalent to the above  $X \rightarrow$  trial at which the  $r$ th success occurs, since  $Y = X - r$ . And the alternative form of the Pascal distribution is

$$\begin{aligned} p_Y(y) = P(Y = y) &= \binom{r+y-1}{y} p^r (1-p)^y \\ &= \binom{-r}{y} p^r (p-1)^y. \end{aligned}$$

As we know negative binomial expansion follows:

$$(1+t)^{-r} = \sum_k \binom{-r}{k} t^k = \sum_k (-1)^k \binom{r+k-1}{k} t^k.$$

The problem of Banach's match boxes [Feller, 1968]. A certain mathematician always carries one match box in his right pocket and one in his left. When he wants a match, he selects a pocket at random, the successive choices thus constituting Bernoulli trials with  $p = 1/2$ . Suppose that initially each box contained exactly  $N$  matches and consider the moment when, for the first time, our mathematician discovers that a box is empty. At that moment the other box may contain  $0, 1, 2, \dots, N$  matches, and we denote the corresponding probabilities by  $u_r$ . Let us identify success with choice of the left pocket. The left pocket will be

found empty at a moment when the right pocket contains exactly  $r$  matches if, and only if, exactly  $N - r$  failures precede the  $(N + 1)$ st success. Therefore the required probability is

$$u_r = \binom{2N - r}{N} \left(\frac{1}{2}\right)^N \left(\frac{1}{2}\right)^{N-r}.$$

Generalization of table tennis [Feller]. Suppose that Peter and Paul play a game which may be treated as a sequence of Bernoulli trials in which the probabilities  $p$  and  $q$  serve as measures for the players' skill. In ordinary table tennis the player who first accumulates 21 individual victories wins the whole game. For comparison with the preceding example we consider the general situation where  $2\nu + 1$  individual successes are required. The game lasts at least  $2\nu + 1$  and at most  $4\nu + 1$  trials. Denote by  $a_r$  the probability that Peter wins at the trial number  $4\nu + 1 - r$ . This event occurs if, and only if, in the first  $4\nu - r$  trials Peter has scored  $2\nu$  successes and thereafter wins the  $(2\nu + 1)$ st trial. So, the required probability is

$$a_r = \binom{4\nu - r}{2\nu} p^{2\nu+1} q^{2\nu-r}.$$

**Example 3.7** (Multinomial distribution). The binomial distribution can easily be generalized to the case of  $n$  repeated independent trials where each trial can have one of several outcomes. Denote the possible outcomes of each trial by  $E_1, E_2, \dots, E_r$ , and suppose that the probability of the realization of  $E_i$  in each trial is  $p_i$  subject to the condition

$$p_1 + p_2 + \dots + p_r = 1.$$

The result of  $n$  trials is a succession like  $E_3 E_1 E_2 \dots$ . The probability that in  $n$  trials  $E_1$  occurs  $k_1$  times,  $E_2$  occurs  $k_2$  times, etc., is

$$p_X(k) = P(X = k = (k_1, k_2, \dots, k_r)) = \frac{n!}{k_1! k_2! \dots k_r!} p_1^{k_1} p_2^{k_2} \dots p_r^{k_r},$$

where,  $k_1 + k_2 + \dots + k_r = n$ . If  $r = 2$ , then the above multinomial distribution reduces to the binomial distribution with  $p_1 = p, p_2 = 1 - p, k_1 = k, k_2 = n - k$ .

For example: in rolling twelve dice, find the probability of getting each face twice. Here  $E_1, E_2, \dots, E_6$  represent the six faces, all  $k_i$  equal 2, and all  $p_i$  equal  $1/6$ . Therefore, the probability is  $\frac{12!}{2^6}(1/6)^{12} = 0.0034$ .

**Example 3.8** (Uniform discrete distribution). A random variable  $X \in \{a, a + 1, \dots, b\}$  has a uniform discrete distribution if

$$p_X(x) = \begin{cases} \frac{1}{b-a+1} & \text{if } x \in \{a, a + 1, \dots, b\}, \\ 0 & \text{if } x \notin \{a, a + 1, \dots, b\}. \end{cases}$$

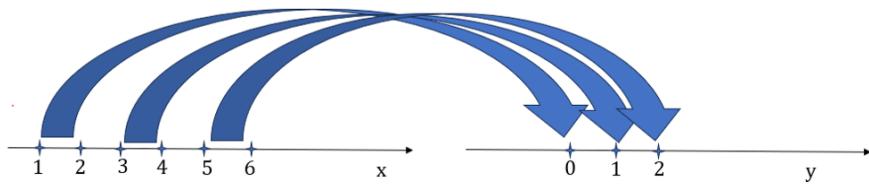
### 3.2.2 Derived distribution of a discrete random variable

Consider a die whose faces are labelled with the numbers 0, 0, 1, 1, 2, 2. We wish to find the PMF of the number observed when the die is tossed, assuming all sides are equally likely to occur. If the original sample space is composed of the possible cube sides that can occur, so that  $\Omega_X = \{1, 2, 3, 4, 5, 6\}$ , then the transformation appears as follows:

$$Y = y = \begin{cases} 0 & \text{if } x = 1 \text{ or } 2 \\ 1 & \text{if } x = 3 \text{ or } 4 \\ 2 & \text{if } x = 5 \text{ or } 6. \end{cases}$$

So, PMF of Y is given by

$$p_Y(y) = \begin{cases} p_X(1) + p_X(2) = 1/3 & \text{if } y = 0, \\ p_X(3) + p_X(4) = 1/3 & \text{if } y = 1, \\ p_X(5) + p_X(6) = 1/3 & \text{if } y = 2. \end{cases}$$



**Figure 3.8:** Transformation or function of a discrete random variable.

If  $Y = g(X)$  is a transformed or derived random variable of a given DRV  $X$ , then the derived distribution  $p_Y(y)$  is given by

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = \sum_{\{x:g(x)=y\}} p_X(x),$$

where  $g^{-1}(y)$  is an inverse image of the observed value  $Y = y$  in  $\Omega_X$ . So, computation of derived distribution  $p_Y(y)$  of a discrete random variable  $X$  follows two steps:

1. Determining all possible observed value of the derived random variable  $Y$  i.e.  $\Omega_Y$ ,
2. Computing the derived distribution as follows:

$$p_Y(y) = \sum_{\{x:g(x)=y\}} p_X(x).$$

**Example 3.9.** If  $X \sim Ber(p)$  and  $Y = 2X - 1$ , then derived distribution  $p_Y(y)$  can be computed as follows:

1. As  $X \sim Ber(p)$  and  $Y = 2X - 1$ , so  $\Omega_X = \{0, 1\}$  transformed to  $\Omega_Y = \{-1, 1\}$ .
2. Now, we have

$$\begin{aligned} p_Y(-1) &= P(Y = -1) = P(2X - 1 = -1) = P(2X = 0) \\ &= P(X = 0) = 1 - p, \\ p_Y(1) &= P(Y = 1) = P(2X - 1 = 1) = P(2X = 2) \\ &= P(X = 1) = p. \end{aligned}$$

**Example 3.10.** If  $X \sim Unif([-4, 4] \cap \mathbb{Z})$  and  $Y = |X|$ , then derived distribution  $p_Y(y)$  can be computed as follows:

1. As  $X \sim Unif([-4, 4] \cap \mathbb{Z})$  and  $Y = |X|$ , so  $\Omega_X = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$  transformed to  $\Omega_Y = \{0, 1, 2, 3, 4\}$ .

2. Now, we have

$$\begin{aligned}
 p_Y(0) &= P(Y = 0) = P(|X| = 0) = P(X = 0) = 1/9, \\
 p_Y(1) &= P(Y = 1) = P(|X| = 1) = P(X \in \{-1, 1\}) = 2/9, \\
 p_Y(2) &= P(Y = 2) = P(|X| = 2) = P(X \in \{-2, 2\}) = 2/9 \\
 p_Y(3) &= P(Y = 3) = P(|X| = 3) = P(X \in \{-3, 3\}) = 2/9 \\
 p_Y(4) &= P(Y = 4) = P(|X| = 4) = P(X \in \{-4, 4\}) = 2/9.
 \end{aligned}$$

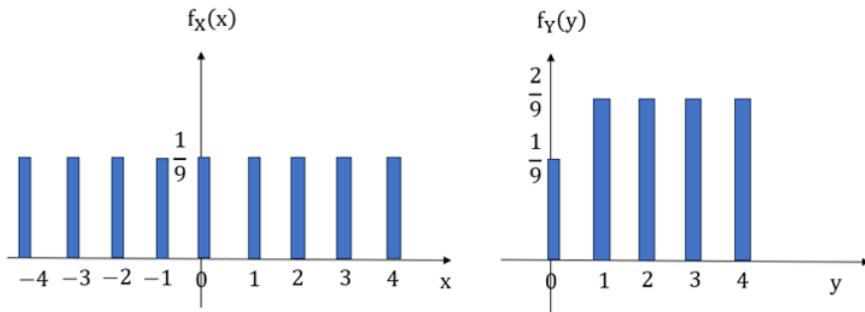


Figure 3.9: Transformation of  $X \sim \text{Unif}([-4, 4] \cap \mathbb{Z})$  to  $Y = |X|$ .

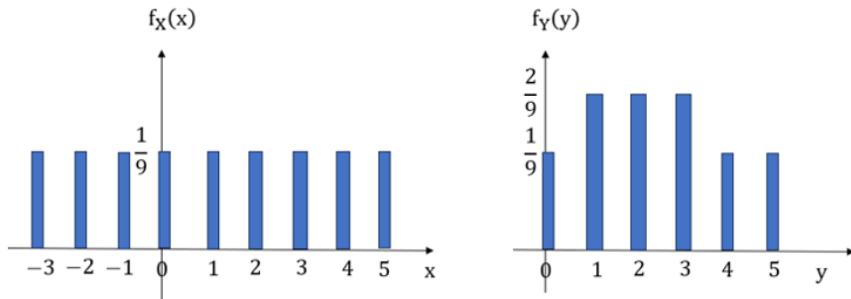
**Example 3.11.** Let  $Y = |X|$  and  $X \sim \text{Unif}([-3, 5] \cap \mathbb{Z})$  is a uniform random variable, then find PMF  $p_Y$  and expectation  $E(Y)$  of Y.

Solution: if  $X \sim \text{Unif}([-4, 4] \cap \mathbb{Z})$  and  $Y = |X|$ , then derived distribution  $p_Y(y)$  can be computed as follows:

- As  $X \sim \text{Unif}([-3, 5] \cap \mathbb{Z})$  and  $Y = |X|$ , so  $\Omega_X = \{-3, -2, -1, 0, 1, 2, 3, 4, 5\}$  transformed to  $\Omega_Y = \{0, 1, 2, 3, 4, 5\}$ .

2. Now, we have

$$\begin{aligned}
 p_Y(0) &= P(Y = 0) = P(|X| = 0) = P(X = 0) = 1/9, \\
 p_Y(1) &= P(Y = 1) = P(|X| = 1) = P(X \in \{-1, 1\}) = 2/9, \\
 p_Y(2) &= P(Y = 2) = P(|X| = 2) = P(X \in \{-2, 2\}) = 2/9 \\
 p_Y(3) &= P(Y = 3) = P(|X| = 3) = P(X \in \{-3, 3\}) = 2/9 \\
 p_Y(4) &= P(Y = 4) = P(|X| = 4) = P(X \in \{4\}) = 1/9 \\
 p_Y(5) &= P(Y = 5) = P(|X| = 5) = P(X \in \{5\}) = 1/9.
 \end{aligned}$$



**Figure 3.10:** Transformation of  $X \sim \text{Unif}([-3, 5] \cap \mathbb{Z})$  to  $Y = |X|$ .

**Example 3.12.** If  $X \sim \text{Unif}([-2, 2] \cap \mathbb{Z})$  and  $Y = X^2$ , then derived distribution  $p_Y(y)$  can be computed as follows:

- As  $X \sim \text{Unif}([-2, 2] \cap \mathbb{Z})$  and  $Y = X^2$ , so  $\Omega_X = \{-2, -1, 0, 1, 2\}$  transformed to  $\Omega_Y = \{0, 1, 4\}$ .

2. Now, we have

$$\begin{aligned}
 p_Y(0) &= P(Y = 0) = P(X^2 = 0) = P(X = 0) = 1/5, \\
 p_Y(1) &= P(Y = 1) = P(X^2 = 1) = P(X \in \{-1, 1\}) = 2/5, \\
 p_Y(4) &= P(Y = 4) = P(X^2 = 4) = P(X \in \{-2, 2\}) = 2/5.
 \end{aligned}$$

**Example 3.13.** If  $X \sim \text{Pois}(\lambda)$  and  $Y = g(X) = \begin{cases} 1 & \text{if } x \text{ is even} \\ -1 & \text{if } x \text{ is odd} \end{cases}$ , then derived distribution  $p_Y(y)$  can be computed as follows:

1. As  $X \sim Pois(\lambda)$  and  $Y = g(X)$ , so  $\Omega_X = \{0, 1, 2, \dots\}$  transformed to  $\Omega_Y = \{-1, 1\}$ .

2. Now, we have

$$\begin{aligned} p_Y(1) &= P(Y = 1) = P(X \text{ is even}) = \sum_{\{x \text{ is even}\}} \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \frac{1}{2} (1 + e^{2\lambda}), \\ p_Y(-1) &= P(Y = -1) = P(X \text{ is odd}) = \sum_{\{x \text{ is odd}\}} \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \frac{1}{2} (1 - e^{-2\lambda}). \end{aligned}$$

### 3.2.3 Jointly discrete distribution

In science and in real life, we are often interested in two (or more) random variables at the same time in a random experiment. In such situations the random variables have a joint distribution that allows us to compute probabilities of events involving both variables and understand the relationship between the variables. This is simplest when the variables are independent. When they are not, we use covariance and correlation as measures of the nature of the dependence between them.

**Definition 3.6** (Joint probability mass function). Suppose X and Y are two discrete random variables and that X takes values  $\{x_1, x_2, \dots, x_m\}$  and Y takes values  $\{y_1, y_2, \dots, y_n\}$ . Then, the ordered pair  $(X, Y)$  take values in the product  $\{(x_1, y_1), (x_1, y_2), \dots, (x_m, y_n)\}$ . So, the joint probability mass function (joint pmf) of X and Y can be defined as the function  $p_{(X,Y)}(x_i, y_j) = P(X = x_i, Y = y_j)$  as the joint probability of the joint outcome  $(X = x_i, Y = y_j)$  and satisfies the following:

1.  $p_{(X,Y)}(x, y) \in [0, 1]$  for any  $(x, y) \in \Omega_X \times \Omega_Y$ ,
2.  $\sum_{\{(x,y) \in \Omega_X \times \Omega_Y\}} p_{(X,Y)}(x, y) = 1$ ,
3.  $P((X, Y) \in B) = \sum_{\{(x,y) \in B\}} p_{(X,Y)}(x, y)$ .

More, explicitly, we can visualize the joint probability mass function for each joint point  $\{(x_i, y_j) | 1 \leq i \leq m, 1 \leq j \leq n\}$  in the table (3.1):

**Table 3.1:** Joint probability table for joint probability mass function  $p_{(X,Y)}$  of  $X \in \Omega_X = \{x_1, x_2, \dots, x_n\}$  and  $Y \in \Omega_Y = \{y_1, y_2, \dots, y_m\}$ .

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_n$
$x_1$	$P_{X,Y}(x_1, y_1)$	$P_{X,Y}(x_1, y_2)$	$\dots$	$P_{X,Y}(x_1, y_j)$	$\dots$	$P_{X,Y}(x_1, y_n)$
$x_2$	$P_{X,Y}(x_2, y_1)$	$P_{X,Y}(x_2, y_2)$	$\dots$	$P_{X,Y}(x_2, y_j)$	$\dots$	$P_{X,Y}(x_2, y_n)$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
$x_i$	$P_{X,Y}(x_i, y_1)$	$P_{X,Y}(x_i, y_2)$	$\dots$	$P_{X,Y}(x_i, y_j)$	$\dots$	$P_{X,Y}(x_i, y_n)$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
$x_m$	$P_{X,Y}(x_m, y_1)$	$P_{X,Y}(x_m, y_2)$	$\dots$	$P_{X,Y}(x_m, y_j)$	$\dots$	$P_{X,Y}(x_m, y_n)$

**Example 3.14.** Question: consider a roll two dice in together. Let  $X$  be the value on the first die and let  $Y$  be the value on the second die. Then both  $X$  and  $Y$  take values 1 to 6 and the joint pmf is  $p(i, j) = 1/36$  for all  $i$  and  $j$  between 1 and 6. Draw the probability table for the joint PMF for each observation of the jointly discrete random variable  $(X, Y)$ . Further, consider an event  $B = Y - X \geq 2$ , then compute its probability.

Answer: It is given that  $X$  be the value on the first die and let  $Y$  be the value on the second die. So, both  $X$  and  $Y$  take values 1 to 6 with joint pmf is  $p(i, j) = 1/36$  for all  $i$  and  $j$  between 1 and 6. Then, the probability table of the joint PMF is given in the table (3.2):

**Table 3.2:** Joint probability table for joint probability mass function  $p_{(X,Y)}$  of  $X \in \Omega_X = \{1, 2, \dots, 6\}$  and  $Y \in \Omega_Y = \{1, 2, \dots, 6\}$ .

$X \setminus Y$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Further, we have

$$\begin{aligned} B &= Y - X \geq 2 \\ &= \{(1, 3), (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 5), (3, 6), (4, 6)\}. \end{aligned}$$

So, the desired probability of the event  $B$  is given by

$$P(B) = P((X, Y) | Y - X \geq 2) = 10 \cdot \frac{1}{36} = \frac{5}{18}.$$

**Example 3.15.** Question: consider a roll two dice in together. Let  $X$  be the value on the first die and let  $T$  be the total on both

dice. Then  $X$  takes values 1 to 6,  $Y$  takes values 2 to 12 and the joint pmf is  $p_{X,T}(i,t) = p_{X,Y}(i, t-i)$  for all  $i \in \{1, 2, 3, 4, 5, 6\}$  and  $t \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Draw a probability table for each possible observation of the jointly discrete random variable  $(X, T)$ .

Answer: It is given that  $X$  be the value on the first die and let  $T$  be the total on both dice. Then  $X$  takes values 1 to 6,  $Y$  takes values 2 to 12 and the joint pmf is  $p_{X,T}(i,t) = p_{X,Y}(i, t-i)$  for all  $i \in \{1, 2, 3, 4, 5, 6\}$  and  $t \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Then, the probability table of the joint PMF is given in the table (3.3):

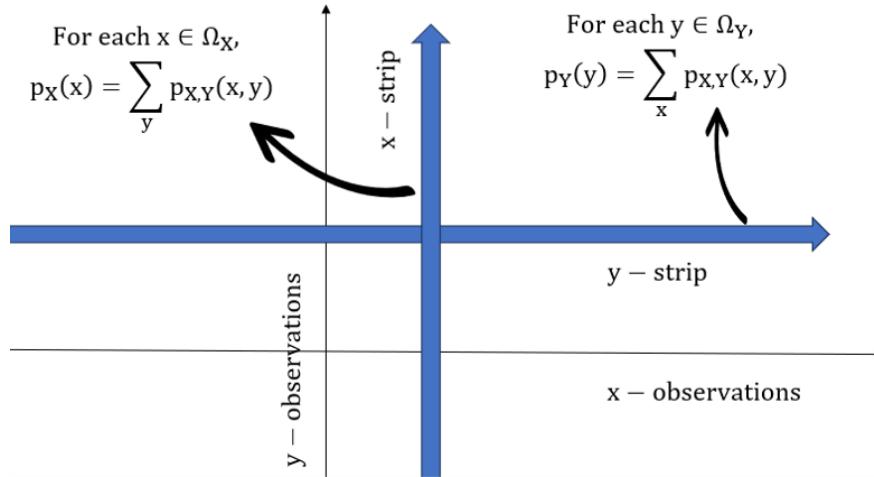
**Table 3.3:** Joint probability table for joint probability mass function  $p_{(X,T)}$  of  $X \in \Omega_X = \{1, 2, \dots, 6\}$  and  $T \in \Omega_T = \{2, 3, \dots, 12\}$ .

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

**Definition 3.7** (Marginal PMFs). Marginalization of joint probability mass function: when  $X$  and  $Y$  are jointly-distributed random variables, we may want to consider only one of them, say  $X$ . In that case we need to find the pmf of  $X$  without  $Y$ . This is called a marginal pmf of  $X$ . For each observation/event  $X = x$ , the  $p_X$  is computed by summing probability of points on  $x$  — strip i.e.

$$p_X(x) = \sum_{y \in \Omega_Y} p_{X,Y}(x,y). \text{ Similarly } p_Y(y) = \sum_{x \in \Omega_X} p_{X,Y}(x,y).$$

**Example 3.16.** Question: consider a roll two dice in together. Let  $X$  be the value on the first die and let  $T$  be the total on both dice. Then  $X$  takes values 1 to 6,  $Y$  takes values 2 to 12 and the joint pmf is  $p_{X,T}(i,t) = p_{X,Y}(i, t-i)$  for all  $i \in \{1, 2, 3, 4, 5, 6\}$  and



**Figure 3.11:** Visualizing marginal PMFs  $p_X(x)$  and  $p_Y(y)$  respectively.

$t \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . In the table each row represents a single value of  $X$ . So the event  $X = 3$  is the third row of the table. To find  $P(X = 3)$  we simply have to sum up the probabilities in this row. We put the sum in the right-hand margin of the table. Likewise  $P(T = 5)$  is just the sum of the column with  $T = 5$ . We put the sum in the bottom margin of the table. Draw a probability table for each possible observation of the jointly discrete random variable  $(X, T)$  along with marginal PMFs.

Answer: It is given that  $X$  be the value on the first die and let  $T$  be the total on both dice. Then  $X$  takes values 1 to 6,  $Y$  takes values 2 to 12 and the joint pmf is  $p_{X,T}(i,t) = p_{X,Y}(i, t-i)$  for all  $i \in \{1, 2, 3, 4, 5, 6\}$  and  $t \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Then, the probability table of the joint PMF is given in the table (3.4):

As motivated by this example, marginal pmf's are obtained from the joint pmf by summing:

$$p_X(x) = \sum_{\{y \in \Omega_Y\}} p_{(X,Y)}(x,y) \quad p_Y(y) = \sum_{\{x \in \Omega_X\}} p_{(X,Y)}(x,y).$$

The term marginal refers to the fact that the values are written in the

**Table 3.4:** Joint probability table for joint probability mass function  $p_{(X,T)}$  of  $X \in \Omega_X = \{1, 2, \dots, 6\}$  and  $T \in \Omega_T = \{2, 3, \dots, 12\}$  along with marginal PMFs  $p_X(x)$  and  $p_T(t)$ .

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12	$p_X(x)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p_T(t)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

margins of the table.

**Definition 3.8** (Conditional probability mass function of a DRV conditioned on an event). In a probabilistic model if a certain event A has already occurred, then we define conditional PMF of the possible observations of a random variable X conditioned on the event A as follows:

$$p_{X|A}(x|A) = P(X = x, A|A) = \frac{P(X = x, A)}{P(A)} = \begin{cases} \frac{p_X(x)}{P(A)} & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

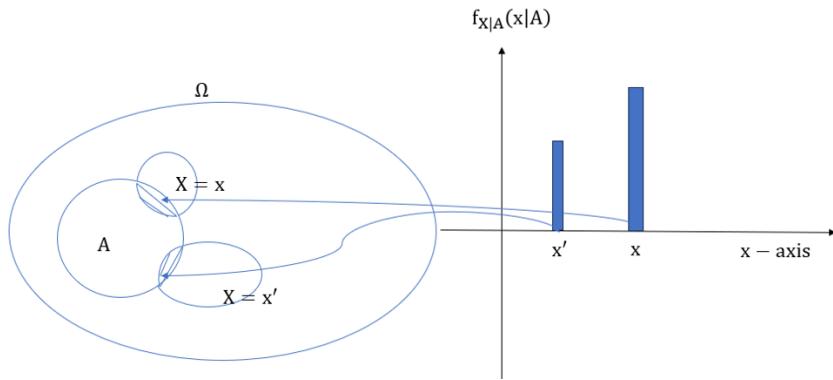
The conditional PMF is a legitimate discrete distribution and satisfies the following:

1.  $p_{X|A}(x|A) \geq 0$  as  $p_X(x) \geq 0$ ,
2. We have the following normalization property:

$$\begin{aligned} \sum_{\{x \in A\}} p_{X|A}(x|A) &= \frac{1}{P(A)} \sum_{\{x \in A\}} p_X(x) = \frac{1}{P(A)} \sum_{\{x \in A\}} P(X = x, A) \\ &= \frac{1}{P(A)} P(A) = 1. \end{aligned}$$

Since for any discrete random variable  $X$ , we have

- a.  $A_i \cap A_j = X^{-1}(x_i) \cap X^{-1}(x_j) = \{\omega | X(\omega) = x_i\} \cap \{\omega | X(\omega) = x_j\}$   
 $= \varphi$  for  $i \neq j$ ,
- b.  $\cup_{\{x_k \in \Omega_X\}} A_k = \cup_{\{x_k \in \Omega_X\}} X^{-1}(x_k) = \cup_{\{x_k \in \Omega_X\}} \{\omega | X(\omega) = x_k\}$   
 $= \Omega$ ,
- c.  $A_k = X^{-1}(x_k) = \{\omega | X(\omega) = x_k\} \neq \varphi$  for any  $x_k \in \Omega_X$ .



**Figure 3.12:** Conditional probability mass function of  $X$  conditioned on an event  $A$ .

**Example 3.17.** Consider an experiment of rolling a dice,  $A$  be the event that roll is an even number. Let  $X$  be the roll of a die, then conditional PMF of  $X$  conditioned  $A$  is given by

$$\begin{aligned} p_{X|A}(x|A) &= P(X = x|A) = \frac{P(X = x, A)}{P(A)} = \begin{cases} \frac{p_X(x)}{P(A)} & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \\ &= \begin{cases} 1/3 & \text{if } x \in A = \{2, 4, 6\} \\ 0 & \text{if } x \notin A. \end{cases} \end{aligned}$$

### 3.2.4 Conditional PMF

**Definition 3.9** (Conditional probability mass function). When an experiment is a compound one, in which the second part of the experiment (described by a discrete random variable  $Y$ ) depends upon the outcome of the first part (described another discrete random variable  $X$ ), conditional probability mass function, conditioned on the event  $Y = y$ , is coming into picture as follows:

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x, Y = y|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{(X,Y)(x,y)}}{p_Y(y)} \\ &= \frac{p_{X,Y}(x,y)}{\sum_x p_{X,Y}(x,y)} \\ &\text{for a fixed observation } Y = y. \end{aligned}$$

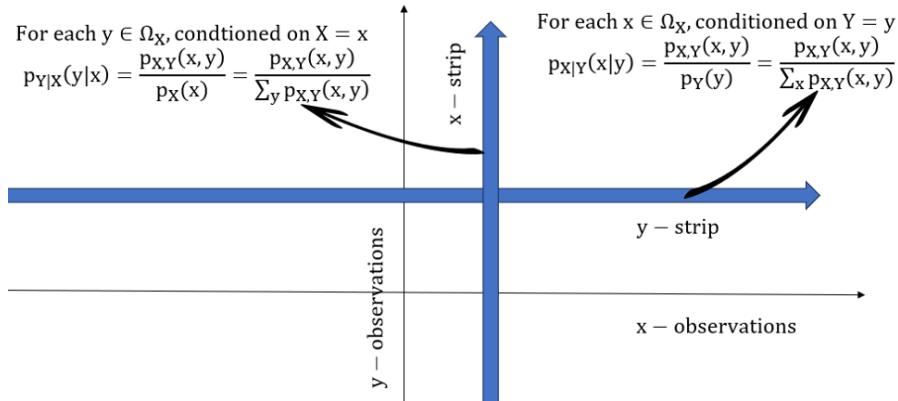
**Definition 3.10** (Conditional PMFs). Conditioning of joint probability mass function: when  $X$  and  $Y$  are jointly-distributed random variables, we may want to consider only one of them conditioned on one observation the other, say  $X|Y = y$ . In that case we need to find the pmf of  $X$  conditioned on  $Y = y$ . This is called a conditional pmf of  $X$  given  $Y = y$ . For each observation/event  $X = x$ , the  $p_{X|Y}$  is computed by a normalized probability of a point on  $x \rightarrow$  strip, where the normalization is provided by the value of PMF  $p_Y$  at  $Y = y$  i.e.  $p_Y(y) = P(Y = y) = \sum_x p_{X,Y}(x,y)$  as a total probability of the event  $Y = y$  on  $y-$  axis. So, we can have a concrete definition of the conditional PMF as follows:

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{(X,Y)(x,y)}}{p_Y(y)} \\ &= \frac{p_{X,Y}(x,y)}{\sum_x p_{X,Y}(x,y)} \\ &\text{for a fixed observation } Y = y. \end{aligned}$$

Similarly,

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

for a fixed observation  $X = x$ .



**Figure 3.13:** Visualizing conditional PMFs  $p_{Y|X}(y|x)$  and  $p_{X|Y}(x|y)$  respectively.

So, by using the definition of conditional probability for events we have

$$\begin{aligned}
 p_{(X,Y)}((x,y)) &= P(X = x, Y = y) = P(Y = y)P(X = x|Y = y) \\
 &= p_Y(y)p_{X|Y}(x|y).
 \end{aligned}$$

It provides a computational framework to determine joint PMF of the joint discrete random variable  $(X, Y)$ .

**Example 3.18.** Two dice are tossed with all outcomes assumed to be equally likely. The number of dots observed on each die are added together. Then compute the conditional PMF of the sum if it is known that the sum is even.

Let  $Y$  be the sum and define  $X = 1$  if the sum is even and  $X = 0$  if the sum is odd. Thus, we wish to determine  $p_{Y|X}(y|0)$  and  $p_{Y|X}(y|1)$  for all  $y \in \Omega_Y = \{2, 3, \dots, 12\}$ . Then

$$p_{Y|X}(y|1) = \frac{p_{(Y,X)}(y, 1)}{p_X(1)} = \begin{cases} 2 \cdot \frac{1}{36} & \text{if } y = 2, 12 \\ 2 \cdot \frac{3}{36} & \text{if } y = 4, 10 \\ 2 \cdot \frac{5}{36} & \text{if } y = 6, 8 \end{cases} = \begin{cases} \frac{1}{18} & \text{if } y = 2, 12 \\ \frac{3}{18} & \text{if } y = 4, 10 \\ \frac{5}{18} & \text{if } y = 6, 8. \end{cases}$$

And

$$p_{Y|X}(y|0) = \frac{p_{(Y,X)}(y, 0)}{p_X(0)} = \begin{cases} 2 \cdot \frac{2}{36} & \text{if } y = 3, 11 \\ 2 \cdot \frac{4}{36} & \text{if } y = 5, 9 \\ 2 \cdot \frac{6}{36} & \text{if } y = 7 \end{cases} = \begin{cases} \frac{2}{18} & \text{if } y = 3, 11 \\ \frac{4}{18} & \text{if } y = 5, 9 \\ \frac{6}{18} & \text{if } y = 7. \end{cases}$$

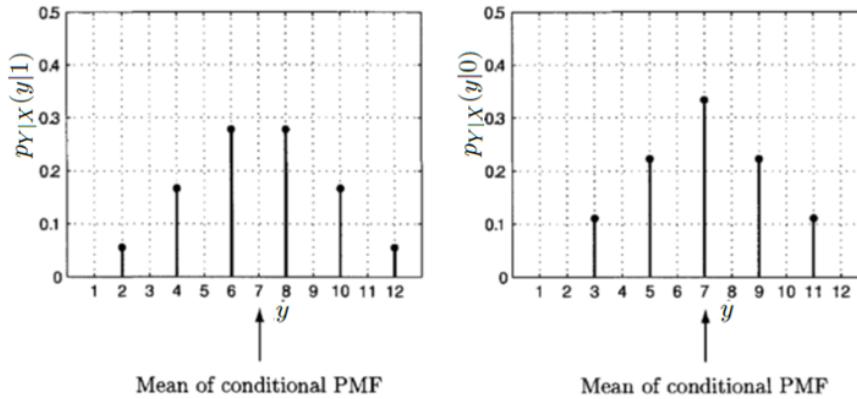
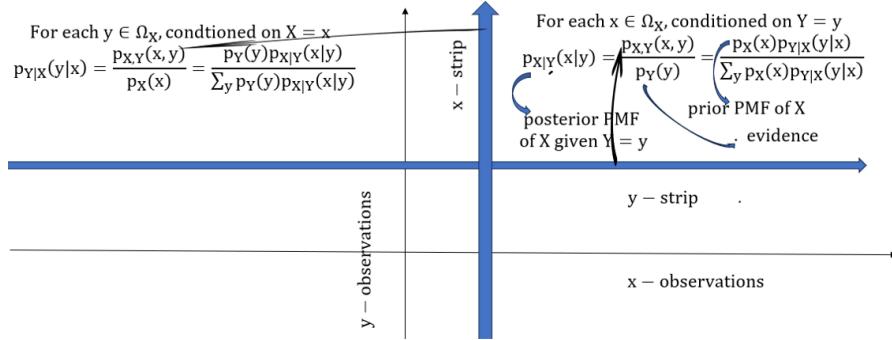


Figure 3.14: Conditional PMFs  $p_{Y|X}(y|1)$  and  $p_{Y|X}(y|0)$ .

**Theorem 3.1** (Bayes rule for probability mass function). Definition of conditional probability mass function leads to computation of posterior distribution as follows:

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

where  $p_X(x) \rightarrow$  prior distribution of  $X$ ,  $p_{Y|X}(y|x) = f(x) \rightarrow$  is the data generating process as a likelihood of  $Y = y$  conditioned on  $X = x$



**Figure 3.15:** Bayes theorem for posterior PMF of  $Y$  conditioned on  $X = x$  and of  $X$  conditioned on  $Y = y$  respectively.

and  $p_Y(y) = \sum_x p_X(x)p_{Y|X}(y|x) \rightarrow$  probability of observing  $Y = y$  or evidence.

**Example 3.19** (Computation of joint PMF). (3.19) Professor May 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  $1/3$ . Find the desired joint probability mass function.

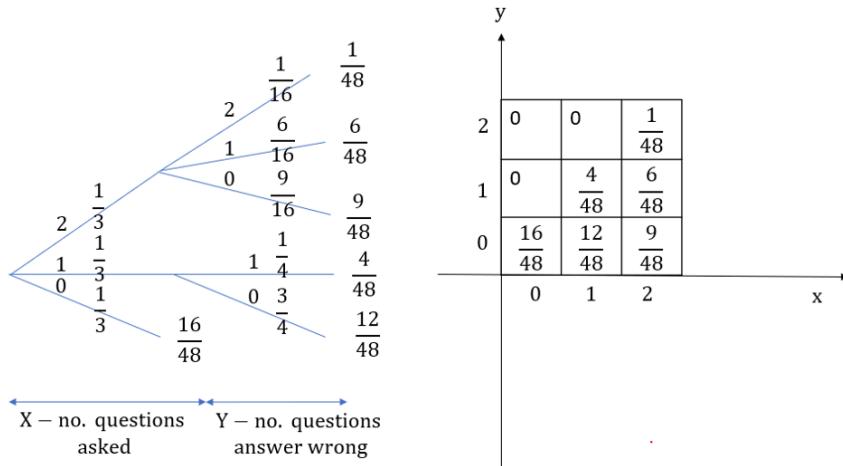
Solution: Let  $X$  and  $Y$  be the number of questions May is asked and the number of questions she answers wrong in a given lecture, respectively. Now need to compute the joint PMF of  $X$  and  $Y$  as follows:

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

We have (a.)  $Y|X = 0 \in \{0\}$  with PMF  $p_{Y|X}(y|0) = 1$ , (b.)  $Y|X = 1 \in \{0, 1\}$  with PMF  $p_{Y|X}(y|0) = \left(\frac{1}{4}\right)^y \left(\frac{3}{4}\right)^{1-y}$ , and (c.)  $Y|X = 2 \in$

$\{0, 1, 2\}$  with PMF  $p_{Y|X}(y|x) = \binom{2}{y} \left(\frac{1}{4}\right)^y \left(\frac{3}{4}\right)^{2-y}$ . Then

$$\begin{aligned} p_{(X,Y)}(0,0) &= p_X(0)p_{Y|X}(0|0) = 1/3 \cdot 1 = 1/3 = 16/48, \\ p_{(X,Y)}(0,1) &= p_X(0)p_{Y|X}(1|0) = 1/3 \times 0 = 0, \\ &\vdots = \vdots \\ p_{(X,Y)}(2,2) &= p_X(2)p_{Y|X}(2|2) = 1/3 \times 1/16 = 1/48. \end{aligned}$$



**Figure 3.16:** Computation of joint PMF of dependent discrete random variables  $X$  and  $Y$ , where  $p_{(X,Y)}(x,y) = p_X(x)p_{Y|X}(y|x)$ .

**Example 3.20** (Computation of posterior PMF). Question: suppose  $X$  is the number of 1s that appear in a binary string of length  $L$ , each bit in the string is equal to zero or one with probability  $1/2$ , and the bits are independent. Suppose that the length of the string is also random and uniformly distributed between 1 and 10. We learn that the binary string contains 4 ones. By using the given information, compute the updated PMF for the length of the string  $L$ .

Answer: Given  $L = \ell$ , we know that  $X$  has the binomial distribution

when  $x \leq \ell$  (the probability is zero otherwise):

$$p_{X|L}(x|\ell) = \begin{cases} \binom{\ell}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{\ell-x} & \text{if } x \leq \ell, \\ 0 & \text{otherwise.} \end{cases}$$

And

$$p_L(\ell) = \begin{cases} \frac{1}{10} & \text{if } \ell = 1, 2, \dots, 10, \\ 0 & \text{otherwise.} \end{cases}$$

Given  $L = \ell$ , we have the binary string contains 4 ones (i.e.  $X = 4$ ), so likelihood of observing  $X = 4$  is given by

$$p_{X|L}(4|\ell) = \binom{\ell}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{\ell-4} = \binom{\ell}{4} \left(\frac{1}{2}\right)^{\ell} = f(\ell).$$

And by suing the prior probability mass function of L, we get

$$p_X(4) = \sum_L p_{(X,L)}(x, \ell) = \sum_L p_L(\ell) p_{X|L}(4|\ell) = \sum_{\ell=1}^{10} \frac{1}{10} \binom{\ell}{4} \left(\frac{1}{2}\right)^{\ell}$$

Then, the updated probability mass function as the posterior PMF of L is given by

$$p_{L|X}(\ell|4) = \frac{p_L(\ell)p_{X|L}(4|\ell)}{p_X(4)} = \begin{cases} \frac{\frac{1}{10} \times \binom{\ell}{4} \left(\frac{1}{2}\right)^{\ell}}{p_X(4)} & \text{if } \ell = 1, 2, \dots, 10, \\ 0 & \text{otherwise.} \end{cases}$$

**Definition 3.11** (**In-dependency or dependency of joint discrete random variables**). Jointly-distributed discrete random variables X and Y are independent if observation of one random variable before the other is going to affect their probability mass functions or their joint PMF is the product of the marginal PMF's:

$$p_{X|Y}(x|y) = p_X(x) \text{ or } p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ or } F_{X,Y}(x,y) = F_X(x)F_Y(y),$$

If  $E(XY) \neq E(X)E(Y)$  or  $Cov(X, Y) \neq 0$ , then the jointly distributed random variables X and Y are dependent.

**Example 3.21.** Roll two dice. Let  $X$  be the value on the first die and let  $Y$  be the value on the second die. Independent discrete variables means the probability in a cell must be the product of the marginal probabilities of its row and column. In the probability table below this is true: every marginal probability is  $1/6$  and every cell contains  $1/36$ , i.e. the product of the marginals and  $X, Y \sim \text{Unif}(\{1, 2, 3, 4, 5, 6\})$ . Therefore  $X$  and  $Y$  are independent.

Question: We rolled two dice and let  $X$  be the value on the first die and  $T$  be the total on both dice. First construct a probability table for joint probability mass function and then establish whether  $X$  and  $Y$  are independent or nor.

Answer: we rolled two dice and let  $X$  be the value on the first die and let  $Y$  be the value on the second die. In the probability table (3.5), the probability in a cell is the product of the marginal probabilities of its row and column. Therefore  $X$  and  $Y$  are independent.

**Table 3.5:** Joint probability table for joint probability mass function  $p_{(X,Y)}$  of  $X \in \Omega_X = \{1, 2, \dots, 6\}$  and  $Y \in \Omega_Y = \{1, 2, \dots, 6\}$  along with marginal PMFs  $p_X(x)$  and  $p_Y(y)$ .

$X \setminus Y$	1	2	3	4	5	6	$p_X(x)$
1	$\frac{1}{36}$						
2	$\frac{1}{36}$						
3	$\frac{1}{36}$						
4	$\frac{1}{36}$						
5	$\frac{1}{36}$						
6	$\frac{1}{36}$						
$p_Y(y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

**Example 3.22.** Question: We rolled two dice and let  $X$  be the value on the first die and  $T$  be the total on both dice. First construct a

probability table for joint probability mass function and then establish whether X and T are independent or nor.

Answer: we rolled two dice and let X be the value on the first die and T be the total on both dice. In the probability table (3.6), below most of the cell probabilities are not the product of the marginal probabilities. For example, none of marginal probabilities are 0, so none of the cells with 0 probability can be the product of the marginals.

**Table 3.6:** Joint probability table for joint probability mass function  $p_{(X,T)}$  of  $X \in \Omega_X = \{1, 2, \dots, 6\}$  and  $T \in \Omega_T = \{2, 3, \dots, 12\}$  along with marginal PMFs  $p_X(x)$  and  $p_T(t)$ .

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12	$p_X(x)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p_T(t)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

**Example 3.23.** Consider a tossing of a biased coin (with probability of success is  $1/4$ ) in the following 2 steps: in the first step we flip the coin until we get a heads and let X denote the trial on which the first heads occurs. In the second step we flip the coin X more times and let Y be the number of heads in the second step. Compute joint PMF of X and Y.

Here, the random variable X is geometrically distributed with parameter  $p = 1/4$ , so

$$p_X(x) = p(1-p)^{x-1} = \left(\frac{3}{4}\right)^{x-1} \frac{1}{4} \text{ for } x = 1, 2, \dots$$

And, the conditional PMF of Y conditioned on  $X = x$  is given by

$$p_{Y|X}(y|x) = \binom{x}{y} \left(\frac{1}{4}\right)^y \left(\frac{3}{4}\right)^{x-y} \text{ for } y = 0, 1, 2, \dots, x.$$

And hence, the joint PMF can be computed as follows:

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = \binom{x}{y} \left(\frac{1}{4}\right)^y \left(\frac{3}{4}\right)^{x-y} \left(\frac{3}{4}\right)^{x-1} \frac{1}{4}$$

### 3.2.5 Expectation and variance of discrete random variable

**Definition 3.12** (Expectation of a discrete random variable). Consider a discrete random variable, if we observe  $n$  random values of  $X$ , then the mean of the  $n$  values will be approximately equal to expectation of  $X$ ,  $E(X)$ , for large  $n$ . The expectation,  $E(X)$  is defined as follows:

$$E(X) = \sum_{\{x \in \Omega_X\}} xp_X(x).$$

**Definition 3.13** (Expected value rule). Consider a discrete random variable and let  $Y = g(X)$  be a function of  $X$ . If we observe  $X$  many times ( $n$  times) to give results  $x_1, x_2, \dots, x_n$  and hence  $g$  provides the observations  $g(x_1), g(x_2), \dots, g(x_n)$ . The mean of  $g(x_1), g(x_2), \dots, g(x_n)$  approaches  $E(g(X))$ , for large  $n$ . The expectation,  $E(g(X))$  is defined as follows:

$$\begin{aligned} E(g(X)) &= E(Y) = \sum_{\{y \in \Omega_Y\}} y p_Y(y) = \sum_{\{y \in \Omega_Y\}} y \sum_{g^{-1}(y)=\{x|g(x)=y\}} p_X(x) \\ &= \sum_{\{x \in \Omega_X\}} g(x) p_X(x). \end{aligned}$$

The expectation tells us important information about the average value of a random variable, but there is a lot of information that it doesn't provide. If you are investing in the stock market, the expected rate of return of some stock is not the only quantity you will be interested in, you would also like to get some idea of the riskiness of the investment. The typical size of the fluctuations of a random variable around its expectation is described by another summary statistic, the variance.

**Definition 3.14** (Variance of a discrete random variable). For a discrete random variable  $X$ , the variance,  $Var(X)$ , of  $X$  is defined by

$$Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2.$$

provided that this quantity exists. We can also treat variance,  $Var(X)$ , of  $X$  as an expected value  $E(g(X))$  of the function  $g(X) = (X - E(X))^2$ . That is variance is a measure of how much the distribution of  $X$  is spread out about its mean and only randomness gives rise to variance. Statisticians often prefer to use the standard deviation (in practice for interpretation and visualization) rather than variance (in principle for theoretical analysis and understanding) as a measure of spread.

**Definition 3.15** (Covariance of two a joint discrete random variable). Consider a joint discrete random variable  $(X, Y)$ , where  $X$  and  $Y$  are associated with the same random experiment, then covariance of  $(X, Y)$  is defined as follows:

$$\begin{aligned} Cov(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - XE(Y) - YE(X) + E(X)E(Y)) = E(XY) - E(X)E(Y). \end{aligned}$$

**Theorem 3.2.** If  $X$  and  $Y$  are independent discrete random variables whose expectations exist, then

$$(a.) E(XY) = E(X)E(Y), \quad (b.) E(g(X)h(Y)) = E(g(X))E(h(Y)).$$

**Example 3.24** (A counter-example for converse of (3.18)). Consider a uniformly distributed discrete random variable,  $X \in \{-1, 0, 1\}$  i.e.  $X \sim Unif([-1, 1] \cap \mathbb{Z})$ , then

$$\begin{aligned} a. \quad &E(X) = 0, \quad b. \quad E(X^2) = \frac{2}{3}, \quad c. \quad E(X^3) = 0, \\ d. \quad &Cov(X, X^2) = E(X^3) - E(X)E(X^2) = 0. \end{aligned}$$

For the observation  $X^2 = t$ ,

$$p_{X|X^2}(x|t) = \frac{P(X = x, X^2 = t)}{P(X^2 = t)} = \frac{P(X = x, X \in \{-\sqrt{t}, \sqrt{t}\})}{P(X \in \{-\sqrt{t}, \sqrt{t}\})}.$$

So,

$$p_{X|X^2}(\sqrt{t}|t) = \frac{p_X(\sqrt{t})}{p_X(-\sqrt{t}) + p_X(\sqrt{t})} \neq p_X(\sqrt{t}).$$

Also,

$$p_{X|X^2}(-\sqrt{t}|t) = \frac{p_X(-\sqrt{t})}{p_X(-\sqrt{t}) + p_X(\sqrt{t})} \neq p_X(-\sqrt{t})$$

Therefore  $X$  and  $X^2$  are dependent discrete random variable despite of having zero covariance.

**Theorem 3.3.** If  $X$  is a non-negative and non-constant discrete random variable, then  $X$  and  $X^2$  are positively correlated, and hence dependent.

*Proof.* We have  $X \geq 0$  and  $X^2 \geq 0$ , then

$$\begin{aligned} Cov(X, X^2) &= E((X - E(X))(X^2 - E(X^2))) \\ &= E((X - E(X))(X^2 - (E(X))^2 + (E(X))^2 - E(X^2))) \\ &= E((X - E(X))(X^2 - (E(X))^2)) \\ &\quad + E((X - E(X))((E(X))^2 - E(X^2))) \\ &= E((X - E(X))(X^2 - (E(X))^2)) \\ &= E((X - E(X))^2(X + E(X))) > 0. \end{aligned}$$

□

**Example 3.25.** There are 5 candidates for 2 job openings, 3 of the candidates are women and 2 are men. Let  $X$  be the number of women hired. Find PMF, expected value, and variance of  $X$ .

Solution: the count of total possible outcomes of hiring 2 person from 5 candidates as a size of the sample is given by

$$|\Omega| = \binom{5}{2} = 10.$$

Suppose,  $X$  be the number of women hired, then probability mass function of  $X$  is computed as:

$$\begin{aligned} \{X = 0\} \implies |\{X = 0\}| &= \binom{3}{0} \cdot \binom{2}{2} = 1 \implies p_X(0) = P(X = 0) = \frac{1}{10}. \\ \{X = 1\} \implies |\{X = 1\}| &= \binom{3}{1} \cdot \binom{2}{1} = 6 \implies p_X(1) = P(X = 1) = \frac{6}{10}. \\ \{X = 2\} \implies |\{X = 2\}| &= \binom{3}{2} \cdot \binom{2}{0} = 3 \implies p_X(2) = P(X = 2) = \frac{3}{10}. \end{aligned}$$

Then

$$E(X) = \sum_{x=0}^2 xp_X(x) = \frac{12}{10}, \quad E(X^2) = \sum_{x=0}^2 x^2 p_X(x) = \frac{18}{10} \implies Var(X) = \frac{36}{100}.$$

**Example 3.26** (Expectation and variance of various discrete distributions).

Determine the expectation and variance of the following discrete random variables:

(a.)  $X \sim Ber(p)$ ,

**Solution 3.1.** As  $X \sim Ber(p)$ , so  $p_X(x) = p^x(1-p)^{1-x}$  for  $x \in \{0, 1\}$ . Then

$$E(X) = \sum_{x \in \{0,1\}} xp_X(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

And,

$$E(X^2) = \sum_{x \in \{0,1\}} x^2 p_X(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Then,

$$Var(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p).$$

(b.)  $Y \sim Bin(n, p)$ ,

**Solution 3.2.** As  $Y \sim Ber(p)$ ,  $p_Y(y) = \binom{n}{y} p^y(1-p)^{n-y}$  for  $y \in \{0, 1, 2, \dots, n\}$ . Then

$$\begin{aligned} E(Y) &= \sum_{y \in \{0,1,\dots,n\}} y p_Y(y) \\ &= \sum_{y \in \{0,1,\dots,n\}} y \binom{n}{y} p^y (1-p)^{n-y} \\ &= \sum_{y \in \{0,1,\dots,n\}} n \binom{n-1}{y-1} p^y (1-p)^{n-y} = np. \end{aligned}$$

And,

$$\begin{aligned}
E(Y^2) &= \sum_{y \in \{0,1,\dots,n\}} y^2 p_Y(y) \\
&= \sum_{y \in \{0,1,\dots,n\}} y^2 \binom{n}{y} p^y (1-p)^{n-y} \\
&= \sum_{y \in \{0,1,\dots,n\}} npy \binom{n-1}{y-1} p^{y-1} (1-p)^{n-y} \\
&= np((n-1)p + 1) = (np)^2 + np(1-p).
\end{aligned}$$

Then,

$$\begin{aligned}
Var(Y) &= E(Y^2) - (E(Y))^2 = (np)^2 + np(1-p) - (np)^2 \\
&= np(1-p).
\end{aligned}$$

(c.)  $Z \sim Geo(p)$ ,

**Solution 3.3.** As  $Z \sim Geo(p)$ , so  $p_Z(z) = p(1-p)^{z-1}$  for  $z \in \{1, 2, \dots\}$ . Then

$$E(Z) = \sum_{z \in \{1,2,\dots\}} z p_Z(z) = \frac{1}{p}.$$

And,

$$E(Z(Z-1)) = \sum_{z \in \{1,2,\dots\}} z(z-1) p_Z(z) = \frac{2(1-p)}{p^2}.$$

Then,

$$\begin{aligned}
Var(Z) &= E(Z^2) - (E(Z))^2 = E(Z(Z-1)) + E(Z) - (E(Z))^2 \\
&= \frac{1-p}{p^2}.
\end{aligned}$$

(d.)  $W \sim Poiss(\lambda)$ .

**Solution 3.4.** As  $W$ , we have  $p_W(w) = \frac{\lambda^w e^{-\lambda}}{w!}$  for  $w = 0, 1, 2, \dots$  with  $E(W) = \lambda = Var(W)$ .

**Definition 3.16** (Conditional expectation). In a probabilistic model if a certain event A has already occurred, then we define conditional expectation of the possible observations of a random variable X conditioned on the event A as follows:

$$E(X|A) = \sum_{\{x \in \Omega_X\}} xp_{X|A}(x|A).$$

Moreover, if the event A is characterized by another discrete random variable Y in the same experiment as  $Y = y$  and the above summation is absolutely convergent, then the conditional expectation of X conditioned on  $Y = y$  is defined as follows:

$$E(X|Y = y) = \sum_{\{x \in \Omega_X\}} xp_{X|Y}(x|y) = g(y).$$

That is for each possible observation  $Y = y$ ,  $E(X|Y = y) = g(y)$ , and hence conditional expectation defines a random variable (as a function of random variable) as follows:

$$g(Y) = E(X|Y).$$

**Theorem 3.4** (Law of iterated expectation). If  $\{B_1, B_2, \dots\}$  be a partition of  $\Omega$  such that  $P(B_i) > 0 \forall i$ , and X is a discrete random variable in the same experiment, then

$$E(X) = \sum_{i \geq 1} E(X|B_i)P(B_i).$$

Moreover, if Y is a discrete random variable such that  $Y^{-1}(y_i) = B_i = \{\omega | Y(\omega) = y_i\}$ , then

$$\begin{aligned} E(X) &= \sum_{\{y \in \Omega_Y\}} E(X|Y = y)P(Y = y) = \sum_{\{y \in \Omega_Y\}} E(X|Y = y)p_Y(y) \\ &= E(E(X|Y)). \end{aligned}$$

*Proof.* We have

$$\begin{aligned} E(X) &= \sum_{\{x \in \Omega_X\}} xp_X(x) = \sum_x xP(X = x) = \sum_x x \sum_i P(X = x, B_i) \\ &= \sum_x x \sum_i P(B_i)P(X = x|B_i) = \sum_i P(B_i) \sum_x xP(X = x|B_i) \\ &= \sum_i E(X|B_i)P(B_i). \end{aligned}$$

Further, We have

$$\begin{aligned}
 LHS &= E(E(X|Y)) = \sum_{\{y \in \Omega_Y\}} E(X|Y=y)p_Y(y) \\
 &= \sum_{\{y \in \Omega_Y\}} \sum_{\{x \in \Omega_X\}} xp_{X|Y}(x|y)p_Y(y) = \sum_{\{y \in \Omega_Y\}} \sum_{\{x \in \Omega_X\}} xp_{(X,Y)}(x,y) \\
 &= \sum_{\{x \in \Omega_X\}} x \sum_{\{y \in \Omega_Y\}} p_{(X,Y)}(x,y) = \sum_{\{x \in \Omega_X\}} xp_X(x) = E(X) = RHS.
 \end{aligned}$$

□

**Theorem 3.5.** Suppose X and Y are discrete random variables and  $a, b \in \mathbb{R}$  are constants. Then

$$E(aX + bY) = aE(X) + bE(Y)$$

provided that both  $E(X)$  and  $E(Y)$  exist.

*Proof.*

$$\begin{aligned}
 E(aX + bY) &= \sum_x \sum_y (ax + by)p_{(X,Y)}(x,y) = a \sum_x x \sum_y p_{(X,Y)}(x,y) \\
 &\quad + b \sum_y y \sum_x p_{(X,Y)}(x,y) = a \sum_x xp_X(x) + b \sum_y yp_Y(x) \\
 &= aE(X) + bE(Y).
 \end{aligned}$$

□

**Example 3.27.** Let  $X$  be the number of rolls of a fair die required to get the first 6. (So  $X$  is geometrically distributed with parameter  $1/6$ .) Now, computation of expectation,  $E(X)$ , and variance,  $Var(X)$ , of  $X$  as below.

Let  $B_1$  be the event that the first roll of the die gives a 6 (i.e.  $X|B_1 \in \{1\}$  or  $\Omega_{X|B_1} = \{1\}$ , so that  $B_1^c$  is the event that it does not (so  $X|B_1^c \in \{2, 3, \dots\}$  or  $\Omega_{X|B_1^c} = \{2, 3, \dots\}$ ) or number of rolls till first 6 given the first roll is nor 6 or  $1 + X$  as  $\underline{\text{1}} \rightarrow \underline{\text{X}}$ ). Then

$$\begin{aligned}
 E(X) &= E(X|B_1)P(B_1) + E(X|B_1^c)P(B_1^c) = \frac{1}{6} + \frac{5}{6}E(X|B_1^c) \\
 &= \frac{1}{6} + \frac{5}{6}E(X+1) = 1 + \frac{5}{6}E(X) \implies E(X) = 6.
 \end{aligned}$$

Above computation is much simpler as compared a direct calculation using the probability mass function:

$$E(X) = \sum_{x=1}^{\infty} x \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^{x-1} = \frac{1}{1/6} = 6.$$

$$\begin{aligned} E(X^2) &= E(X^2|B_1) + E(X^2|B_1^c)P(B_1^c) = \frac{1}{6} + \frac{5}{6}E(X^2|B_1^c) \\ &= \frac{1}{6} + \frac{5}{6}E((1+X)^2) = \frac{1}{6} + \frac{5}{6}(1 + 2E(X) + E(X^2)) \\ &= 11 + \frac{5}{6}E(X^2) \implies E(X^2) = 66. \end{aligned}$$

And hence variance  $Var(X) = E(X^2) - (E(X))^2 = 66 - 36 = 30$ . Alternatively, we compute the variance as follows:

$$E(X^2) = \sum_{x=1}^{\infty} x^2 \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^{x-1} = 66 \implies Var(X) = 30.$$

**Example 3.28.** Your spaghetti bowl contains  $n$  strands of spaghetti. You repeatedly choose 2 ends at random, and join them together. Compute the average number of loops in the bowl, once no ends remain.

Answer:

- a. On the first try you have  $n$  noodles and  $2n$  ends. You pick an end and the probability that you pick up the other end of the same noodle and make a loop is 1 out of  $2n - 1$  i.e  $\frac{1}{2n-1}$ .
- b. On the next try you have  $n-1$  noodles and  $2n - 2$  ends. You pick an end and the probability that you pick up the other end of the same noodle and make a loop is 1 out of  $2n - 3$  i.e  $\frac{1}{2n-3}$ .
- c. Since, formation of each loop decreases the number of ends by 2. So, on the  $i$ th try you  $(n - i)$  noodles and  $2(n - i) + 1$  ends. You pick an end and the probability that you pick up the other end of the same noodle and make a loop is 1 out of  $2(n - i) + 1$ .

Let  $X_i$  be the indicator function of the event that we form a loop at the  $i$ th step. Then  $E(X_i) = 1/(2(n - i) + 1)$ . Suppose  $T$  be the total

number of loops formed, then  $T = E(X_1 + X_2 + \dots + X_n)$ . So, using linearity of expectation, we have

$$\begin{aligned} E(T) &= E(X_1 + X_2 + \dots + X_n) = \frac{1}{2n-1} + \frac{1}{2n-3} + \dots + 1/3 + 1 \\ &\approx \log(n) \text{ for large } n. \end{aligned}$$

Alternatively, let the expected number of loops from  $n$  noodles be  $E(n)$ . Then, obviously,  $E(1) = 1$ . Now, for  $n > 1$ , if you pick up two ends, the possibility for those two ends belongs to the same noodle will be  $\frac{1}{2n-1}$ . Then you have one loop now and there are  $n - 1$  noodles to keep on going. Otherwise, you get no loop and still have  $n - 1$  noodles to go (the two you pick before are now connected as one). Pick up two ends. They either belong to the same noodle ( $S$ ) or they don't ( $S'$  i.e. no loop addition from the two end). So,

$$\begin{aligned} E(n) &= E(n|S)P(S) + E(n|S')P(S') = (1 + E(n-1)) \cdot \frac{1}{2n-1} \\ &\quad + E(n-1) \cdot \left(1 - \frac{1}{2n-1}\right) = E(n-1) + \frac{1}{2n-1}. \end{aligned}$$

Then, inductively you may expect that

$$E(n) = 1 + \frac{1}{3} + \frac{1}{5} + \dots + \frac{1}{2n-1} \rightarrow \ln(n) \text{ for large } n.$$

### 3.2.6 Function of jointly discrete random variables

The purpose of assessing the student's selection in a campus placement, we might wish to know his/her cgpa, numbers of desired courses, number of projects, etc. All these measurements and others are used by an interview panel/recruiter agency for an offer assessment of a candidate.

**Theorem 3.6** (An another form of total probability). In general to compute probabilities of events it is advantageous to use conditioning arguments whether or not  $X$  and  $Y$  are independent.

$$P(Y \in A) = \sum_x P(Y \in A|X = x) p_X(x) dx = \sum_x \sum_A p_{Y|X}(y|x)p_X(x).$$

**Corollary 3.7** (PDF of a function of jointly discrete random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly discrete and independent

random variables and a function  $Z = \frac{Y}{X}$ , then PMF of  $Z$  is computed as follows:

$$p_Z(z) = \sum_x p_X(x)p_Y(xz).$$

*Proof.* Derivation as follows:

$$\begin{aligned} p_Z(z) &= P(Z = z) = P(Y = zx) = \sum_x P(Y = zx | X = x) p_X(x) \\ &= \sum_x p_{Y|X}(zx|x)p_X(x) = \sum_x p_X(x)p_Y(zx). \end{aligned}$$

□

**Corollary 3.8** (PDF of a function of jointly discrete random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly discrete and independent random variables and a function  $Z = XY$ , then PMF of  $Z$  is computed as follows:

$$p_Z(z) = \sum_x p_X(x)p_Y(z/x).$$

*Proof.* Derivation as follows:

$$\begin{aligned} p_Z(z) &= P(Z = z) = P\left(Y = \frac{z}{X}\right) = \sum_x P\left(Y = \frac{z}{x} | X = x\right) p_X(x)dx \\ &= \sum_x p_X(x)p_{Y|X}(z/x|x) = \sum_x p_X(x)p_Y(z/x). \end{aligned}$$

□

Conditional PMFs can be used to simplify probability calculations. For example, consider the determination of the PMF for  $Z = X + Y$ , where  $X$  and  $Y$  are jointly discrete and independent random variables as  $p_Z(z) = p_X p_Y(z)$ . To solve this problem using conditional PMFs, we ask ourselves the question: Could I find the PMF of  $Z$  if  $X$  were known? If so, then we should be able to use conditioning arguments to first find the conditional PMF of  $Z$  given  $X$ , and then uncondition the result to yield the PMF of  $Z$ .

**Corollary 3.9** (PDF of a function of jointly discrete random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly discrete and independent

random variables and a function  $Z = X + Y$ , then PMF of  $Z$  is computed as follows:

$$p_Z(z) = \sum_x p_X(x)p_Y(z - x).$$

*Proof.* Derivation as follows:

$$\begin{aligned} p_Z(z) &= P(Z = z) = P(X + Y = z) = \sum_x P(Y = z - x | X = x) p_X(x) \\ &= \sum_x p_{Y|X}(z - x | x) p_X(x) = \sum_x p_X(x) p_Y(z - x). \end{aligned}$$

□

### 3.3 Continuous distribution

In the previous section, we discussed discrete random variables and the relevant methods employed to describe them probabilistically. The principal assumption necessary in order to do so is that the range of sample space  $\Omega_X$ , which is the set of all possible outcomes, is finite or at most countably infinite. It followed then that a probability mass function (PMF) could be defined as the probability of each sample point and used to calculate the probability of all possible events (which are subsets of the sample space). Most physical measurements, however, do not produce a discrete set of values but rather a continuum of values such as the rainfall measurement data, maximum temperature measured during the day. The number of possible rainfalls or temperatures take values, in an interval (e.g. [20, 60], is in continuum (i.e. infinite and uncountable). Of course, we could always choose to round off the measurement to the nearest unit so that the possible outcomes would then become  $\{20, 21, \dots, 60\}$ . Then, many valid PMFs could be assigned. But this approach compromises the measurement precision and so is to be avoided if possible. What we are ultimately interested in is the probability of any interval, to do so we must extend our previous approaches to be able to handle this new case. And if we later decide that less precision is warranted, such that the rounding of 20.6 to 21 is acceptable, we will still be able to determine the probability of observing 21. To do so we can regard the rounded temperature of 21 as having

arisen from all temperatures in the interval  $A = [20.5, 21.5]$ . Then,  $P(\text{rounded temperature} = 21) = P(A)$ , so that we have lost nothing by considering a continuum of outcomes.

**Definition 3.17** (Continuous random variable). A measurable function  $X : \Omega \rightarrow \Omega_X \subset \mathbb{R}$  such that  $X^{-1}(B) \in \Sigma_\Omega \forall B \in \mathcal{B}$  such that

(a.)  $X^{-1}(B) = \{\omega | X(\omega) \in B\} \in \Sigma_\Omega$  for every  $B \in \mathcal{B}$ , (b.)  $\Omega_X \subset \mathbb{R}$  is at least continuum in nature.

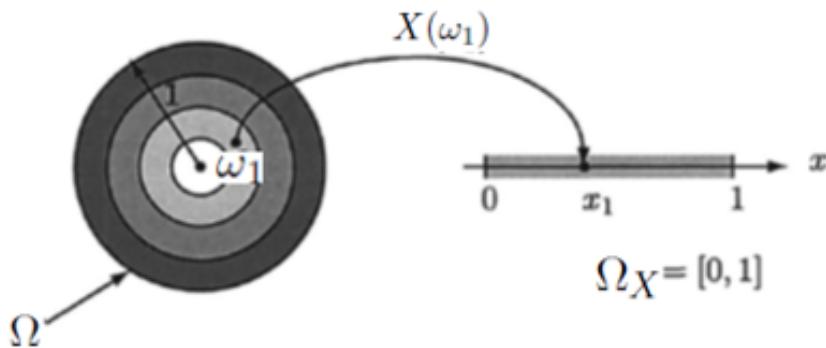
(c.)  $\Omega_X \neq \{x_k\}_{k \in \mathbb{N}} \implies B$  is an interval or union of interval and we bring at most countable scenario using discretization or  $\varepsilon$ - covering.

Then, the measurable function  $X$  is known as a continuous random variable and its probability distribution is characterized by a probability density function. An event of a continuous random variable  $X$  is described as below:

$$X \in B = \{X = x \in [a, b]\} \equiv \{X \in [a, b]\} \equiv \{\omega | X(\omega) \in [a, b]\}$$

for some real numbers  $a, b \in \mathbb{R}$ .

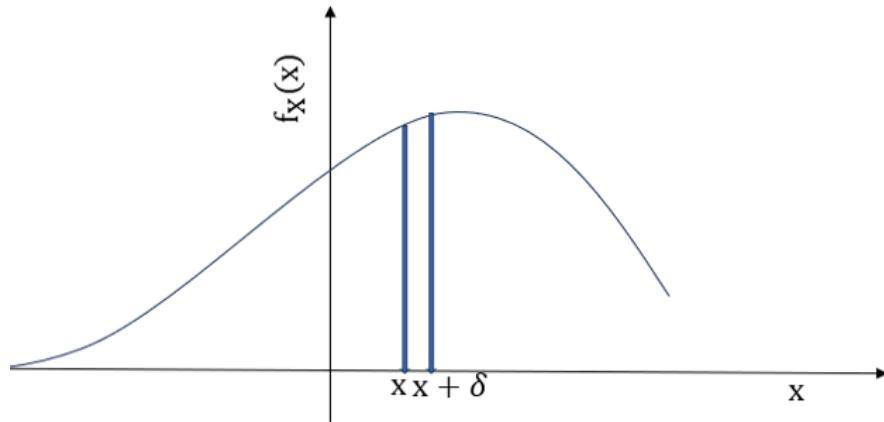
Here,  $\{X \in B\} \rightarrow$  the event that the random variable  $X$  takes the given value  $x$  from an interval  $B = [a, b]$  in continuum manner.



**Figure 3.17:** Continuous random variable as a mapping of an at least continuum sample space into an at least continuum set in  $\mathbb{R}$ .

**Definition 3.18** (Probability density function). Distribution pattern of a continuous random variable is characterized by a probability density function (pdf) (see in (3.18)),  $f : \Omega_X \rightarrow \mathbb{R}_+$  as a probability per unit length and defined as follows:

$$f_X(x) = \lim_{\delta \rightarrow 0} \frac{P(x \leq X \leq x + \delta)}{\delta} \implies P(x \leq X \leq x + \delta) \approx f_X(x)\delta.$$



**Figure 3.18:** Characterization probability density function as a probability per unit length.

The PDF  $p_X$  satisfies the following properties:

- (a.)  $f_X(x) \geq 0$ ,
- (b.)  $\int_{\Omega_X} f_X(x)dx = 1$ ,
- (c.) for any event  $X \in B$ , we have  $P(X \in B) = \int_{x \in B} f_X(x)dx$ .

Nature of an event in term of continuous random variable: if  $X$  is a continuous random variable, then  $\{\omega | X(\omega) \leq x\} = \{X \leq x\} = (-\infty, x]$  is a basic event for every  $x$ . If we have an event  $\{a < X \leq b\}$ , then it is characterized as  $\{a < X \leq b\} = \{X > a\} \cap \{X \leq b\}$ , where  $\{X > a\} = \{X \leq a\}^c$ . Furthermore,  $\{a - \frac{1}{n} < X \leq a + \frac{1}{n}\}$  is an event for every  $n$  and consequently implies  $\bigcup_{n=1}^{\infty} \{a - \frac{1}{n} < X \leq a + \frac{1}{n}\} = \{X = a\}$  is also an event. And, probability of the above events are computed as

follows

$$P(X \in (a, b]) = P(\{a < X \leq b\}) = \lim_{n \rightarrow \infty} \sum_{i=1}^n f_X(x_i) \delta x_i = \int_a^b f_X(x) dx,$$

$$P(\{X = a\}) = \lim_{n \rightarrow \infty} f_X(a) \frac{1}{n} = 0 \text{ as } P(x_{i-1} < X \leq x_i) \approx f_X(a) \delta x_i.$$

### 3.3.1 Various continuous distribution

**Example 3.29** (Uniform distribution). If a continuous random variable  $X$  takes values in an interval  $[a, b]$ , and all subintervals of the same length (with respect to uniform partition of the interval) are equally likely. That is,  $X$  has a uniform continuous distribution (i.e. a uniform PDF) if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b]. \end{cases}$$

So,  $X$  is a uniformly continuous random variable and we write  $X \sim \text{Unif}([a, b])$ . e.g. in a wheel of fortune game, a gambler spins a wheel of fortune, continuously calibrated between 0 and 1, and observes the resulting number. Then, all subintervals of  $[0, 1]$  of the same length (under a uniform discretization scheme) are equally likely, and hence

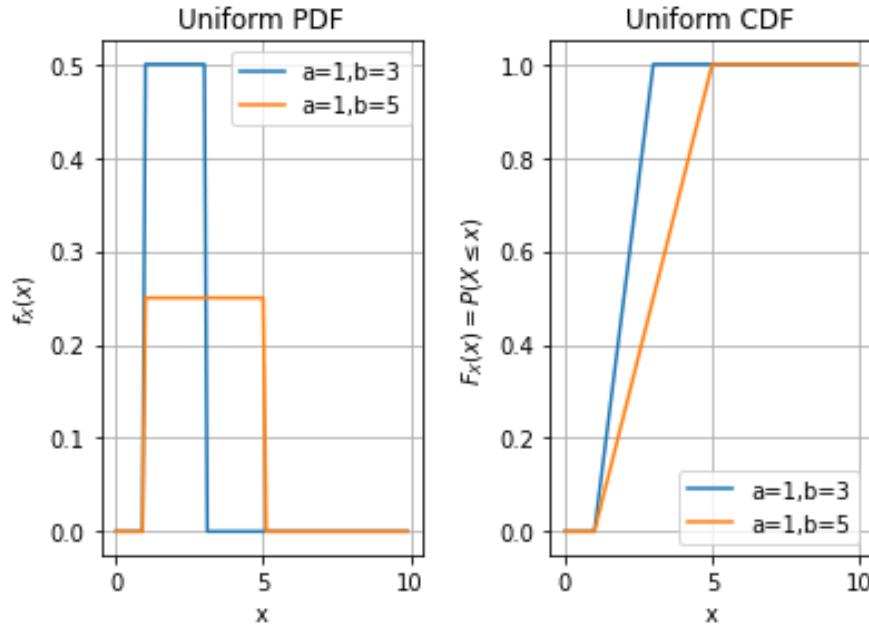
$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{if } x \notin [0, 1]. \end{cases}$$

**Remark 3.3** (Characterization of PDF of continuous random variable). The uniform distribution is the simplest continuous random variable and that obeys uniform law (or equally likely principle) as follows (for a small incremental length  $\delta$ ):

for any two partition points  $x_i, x_j$  of  $[a, b]$ ,

$$\begin{aligned} P(x_i < X \leq x_i + \delta) &= P(x_j < X \leq x_j + \delta) \implies f_X(x_i)\delta = f_X(x_j)\delta \\ \implies f_X(x_i) &= f_X(x_j). \end{aligned}$$

So  $f_X(x) = \frac{1}{b-a} \rightarrow$  is a constant and hence defines a uniform continuous distribution.



**Figure 3.19:** Probability density function as a uniformly distributed random variable on  $[a, b]$ .

For other than uniform continuous random variable, the corresponding PDF is non uniform is characterized as below:

For small values of  $\delta$  and for any two partition points  $x_i, x_j$  of  $[a, b]$ , we have

$$\begin{aligned} P(x_i < X \leq x_i + \delta) &\neq P(x_j < X \leq x_j + \delta) \implies f_X(x_i)\delta \neq f_X(x_j)\delta \\ \implies f_X(x_i) &\neq f_X(x_j). \end{aligned}$$

So  $f_X(x)$  is a non-constant and hence defines a non-uniform continuous distribution.

**Example 3.30** (A piecewise continuous PDF). Alvin's driving time to work is between 15 and 20 minutes if the day is sunny, and between 20 and 25 minutes if the day is rainy, with all times being equally likely in each case. Assume that a day is sunny with probability  $2/3$  and rainy with probability  $1/3$ . Compute the PDF of the driving time.

Here, all times are equally likely in each case means the PDF of  $X$  is constant in each of the two subintervals  $[15, 20]$  and  $[20, 25]$ . And hence

$$f_X(x) = \begin{cases} c_1 & \text{if } x \in [15, 20], \\ c_2 & \text{if } x \in [20, 25], \\ 0 & \text{otherwise.} \end{cases}$$

Further, we have

$$\begin{aligned} \frac{2}{3} &= P(\text{sunny day}) = \int_{15}^{20} f_X(x)dx = 5c_1 \implies c_1 = \frac{2}{15}, \\ \frac{1}{3} &= P(\text{rainy day}) = \int_{20}^{25} f_X(x)dx = 5c_2 \implies c_2 = \frac{1}{15}. \end{aligned}$$

**Example 3.31** (Exponential distribution). A continuous random variable  $X$  takes values in an interval  $[0, \infty)$ , and has an exponential continuous distribution (i.e. an exponential PDF) with a parameter  $\lambda$  if

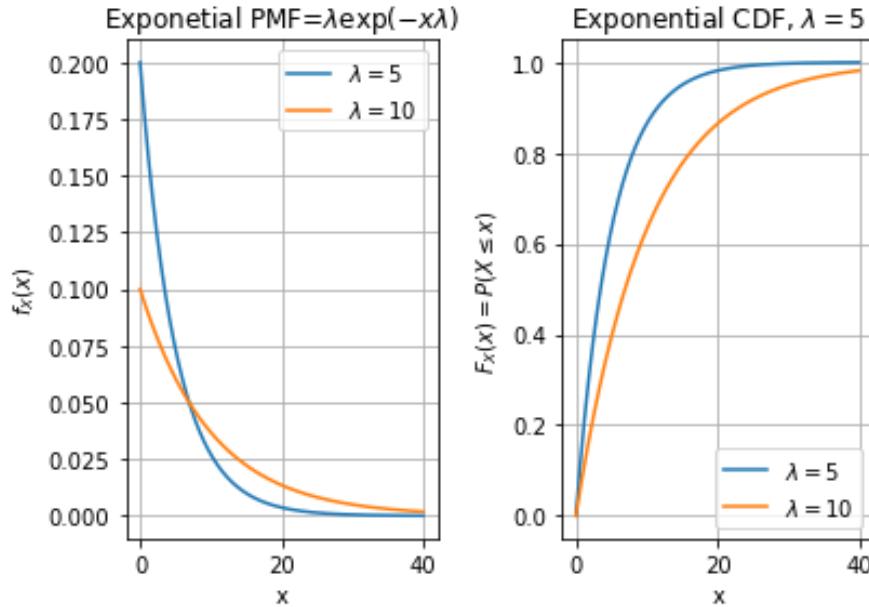
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

So,  $X$  is an exponentially continuous random variable with a discontinuous at  $x = 0$  and we write  $X \sim Exp(\lambda)$ . One important property of exponential distribution is that it provides a formula for right tail probability in order to model life-time or decay-time of a product or object as follows

$$\begin{aligned} P(X > a) &= \int_a^{\infty} f_X(x)dx = \int_a^{\infty} \lambda e^{-\lambda x} dx = \lim_{b \rightarrow \infty} \int_a^b \lambda e^{-\lambda x} dx \\ &= \lim_{b \rightarrow \infty} \left[ -e^{-\lambda x} \right]_a^b = e^{-\lambda a} - \lim_{b \rightarrow \infty} e^{-\lambda b} = e^{-\lambda a}. \end{aligned}$$

For example, if  $X$  is the failure time in days of a light-bulb, then  $P(X > 100)$  is the probability that the light-bulb will fail after 100 days or it will last for at least 100 days. This can be computed as follows:

$$P(X > 100) = e^{-100\lambda} = \begin{cases} 0.367 & \text{if } \lambda = 0.01, \\ 0.904 & \text{if } \lambda = 0.001. \end{cases}$$



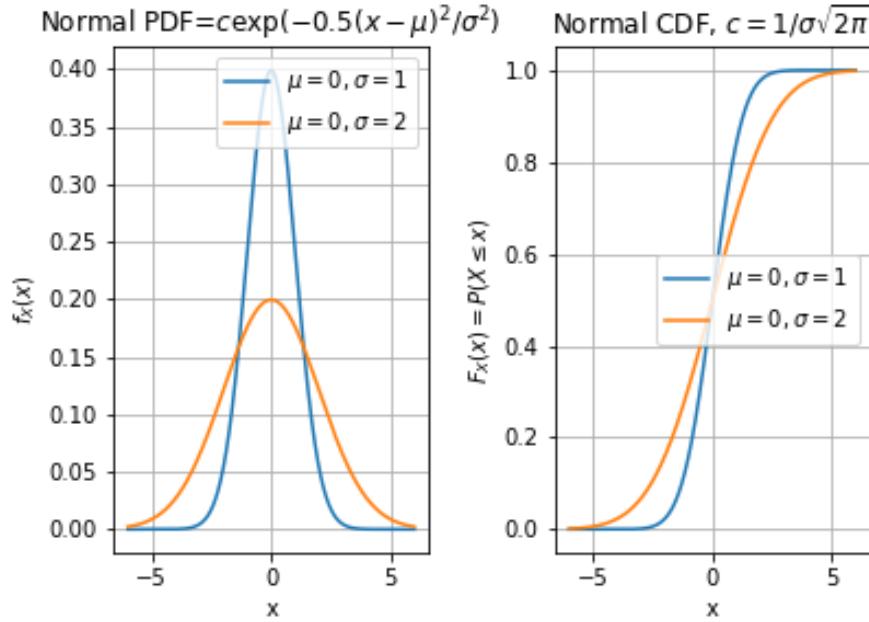
**Figure 3.20:** Probability density function as a exponentially distributed random variable on  $[0, \infty)$ .

**Example 3.32** (Normal distribution). A continuous random variable  $X$  takes values in an interval  $(-\infty, \infty)$ , and has a normal distribution (i.e. a bell-shaped Gaussian PDF) with a pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \text{ for } x \in (-\infty, \infty).$$

So,  $X$  is an normally continuous random variable characterized by the pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  and we write  $X \sim Exp(\lambda)$ .

The parameter  $\mu$  indicates the center of the PDF with a depiction of the average value of the observations made by the random variable  $X$  as a mean or expectation of  $X$ .



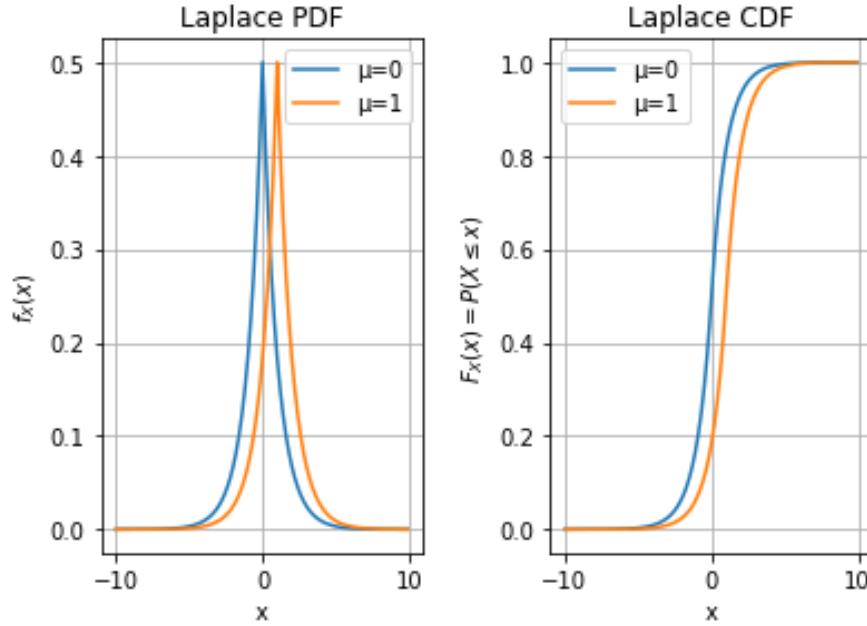
**Figure 3.21:** Probability density function as a normally distributed random variable on  $(-\infty, \infty)$ .

The parameter  $\sigma^2$  indicates the width of the PDF with a depiction of variability of the observations made by the random variable  $X$  as a variance of  $X$ .

**Example 3.33 (Laplace distribution).** A continuous random variable  $X$  takes values in an interval  $(-\infty, \infty)$ , and has a Laplace distribution (similar to normal distribution, but it does not decrease as rapidly from its maximum value) with a pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}|x|\right) \text{ for } x \in (-\infty, \infty).$$

So,  $X$  is an Laplacianally continuous random variable characterized by the pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  and we write



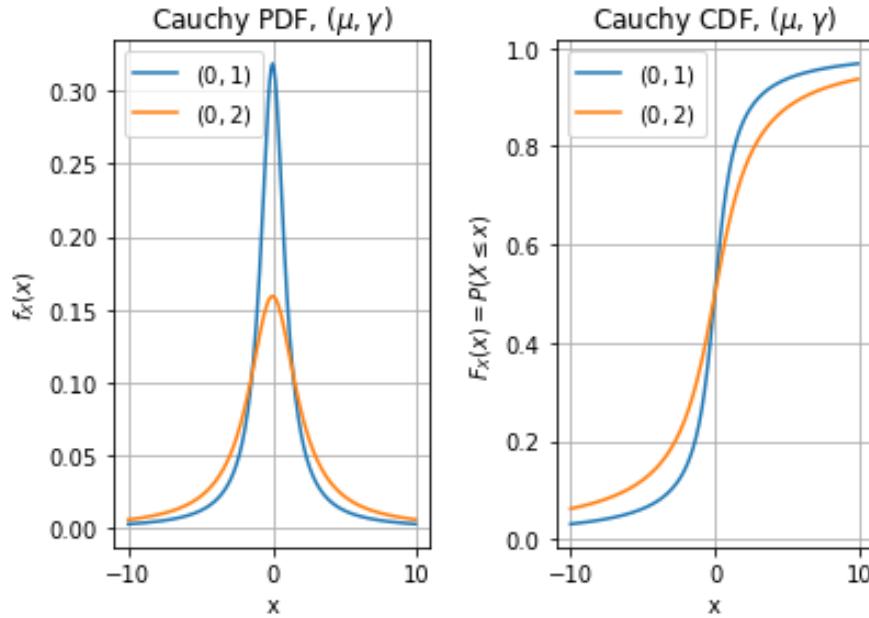
**Figure 3.22:** Probability density function as a Laplace-distributed random variable on  $(-\infty, \infty)$ .

$$X \sim Exp(\lambda).$$

The parameter  $\mu = 0$  indicates the center of the PDF with a depiction of the average value of the observations made by the random variable  $X$  as a mean or expectation of  $X$ .

The parameter  $\sigma^2$  indicates the width of the PDF with a depiction of variability of the observations made by the random variable  $X$  as a variance of  $X$ . Here, the outcomes are larger due to the larger probability in the tails of the PDF (i.e., the tail region of the PDF is that for which  $|x|$  is large). This PDF is used as a model for speech amplitudes.

**Example 3.34 (Cauchy distribution).** A continuous random variable  $X$  takes values in an interval  $(-\infty, \infty)$ , and has a Cauchy distribution (as



**Figure 3.23:** Probability density function as a Cauchy distributed random variable on  $(-\infty, \infty)$ .

a ratio of two standard normal distributions) if

$$f_X(x) = \frac{1}{\pi(1+x^2)} \text{ for } x \in (-\infty, \infty).$$

So,  $X$  is an Cauchy continuous random variable characterized by the pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  and we write  $X \sim Exp(\lambda)$ .

The parameter  $\mu = 0$  indicates the center of the PDF with a depiction of the average value of the observations made by the random variable  $X$  as a mean or expectation of  $X$ .

The parameter  $\sigma^2$  indicates the width of the PDF with a depiction of variability of the observations made by the random variable  $X$  as a variance of  $X$ .

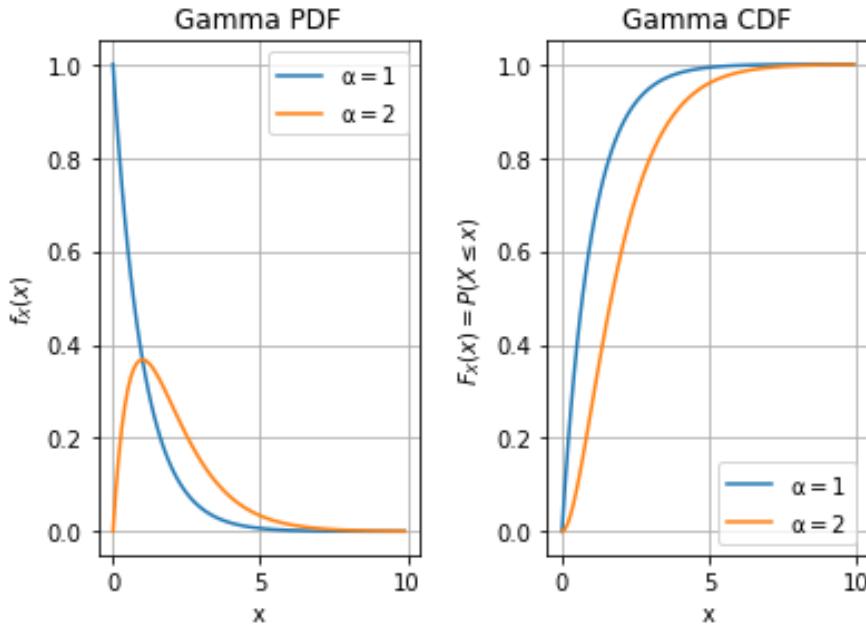
**Example 3.35 (Gamma distribution).** A continuous random variable  $X$  takes values in an interval  $[0, \infty)$ , and has a gamma distribution (as a

generalization of many continuous distributions) if

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

where  $\lambda > 0$ ,  $\alpha > 0$ , and  $\Gamma(\alpha)$  is the Gamma function which is defined as below:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx.$$



**Figure 3.24:** Probability density function as a gamma distributed random variable on  $[0, \infty)$ .

Few properties of the gamma distributed random variable are as follows:

- a.  $\Gamma(\alpha + a) = \alpha\Gamma(\alpha)$ ,

- b.  $\text{Gamma}(n) = (n - 1)!$  for a natural number  $n \in \mathbb{N}$ ,
- c.  $\Gamma(1/2) = \sqrt{\pi}$ .

The Gamma PDF reduces to many well known PDFs for appropriate choices of the parameters  $\alpha$  and  $\lambda$  as follows:

- a. If  $\alpha = 1$ , the gamma distribution reduces to an exponential distribution as follows:

$$f_X(x) = \begin{cases} \frac{\lambda}{\Gamma(1)} \exp(-\lambda x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

- b. If  $\alpha = n/2$  and  $\lambda = 1/2$ , the gamma distribution reduces to a Chi-squared PDF with  $n$  degrees of freedom (i.e. the sum of the squares of  $n$  independent standard normal random variables all with the same PDF  $N(0, 1)$ ) as follows:

$$f_X(x) = \begin{cases} \frac{1}{2^n \Gamma(n/2)} x^{n/2-1} \exp(-x/2) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

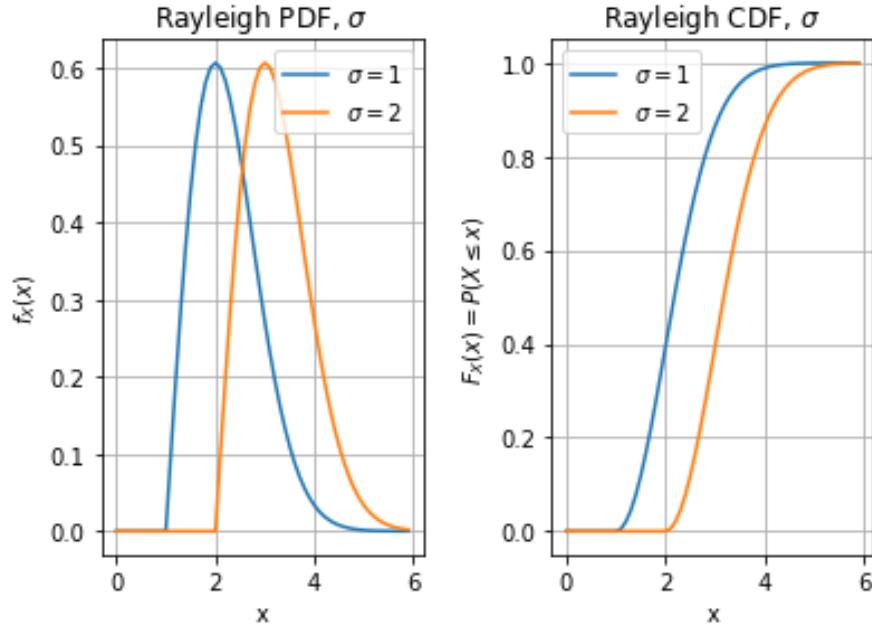
- c. If  $\alpha = n$ , the gamma distribution reduces to an Erlang PDF with  $n$  degrees of freedom (i.e. the sum of  $n$  independent exponential random variables all with the same  $\lambda$ ) as follows:

$$f_X(x) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} x^{n-1} \exp(-\lambda x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

**Example 3.36** (Rayleigh distribution). A continuous random variable  $X$  takes values in an interval  $[0, \infty)$ , and has a Rayleigh distribution (as a square-root of the sum of two independent standard normal distributions with the same PDF  $N(0, 1)$ ) if

$$f_X(x) = \begin{cases} \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

So,  $X$  is an Cauchy continuous random variable characterized by the pair of parameters  $(\mu, \sigma^2)$ ,  $\mu \in (-\infty, \infty)$ ,  $\sigma \in [0, \infty)$  and we write  $X \sim \text{Exp}(\lambda)$ .



**Figure 3.25:** Probability density function of a CRV with Rayleigh distribution on  $[0, \infty)$ .

The parameter  $\mu = 0$  indicates the center of the PDF with a depiction of the average value of the observations made by the random variable  $X$  as a mean or expectation of  $X$ .

The parameter  $\sigma^2$  indicates the width of the PDF with a depiction of variability of the observations made by the random variable  $X$  as a variance of  $X$ .

**Remark 3.4.** Finally, note that many of these PDFs arise as the PDFs of transformed Gaussian random variables. Therefore, realizations of the random variable may be obtained by first generating multiple realizations of independent standard normal or  $N(0, 1)$  random variables, and then performing the appropriate transformation. An alternative and more general approach to generating realizations of a random variable, once the PDF is known, is via the probability integral transformation.

**Example 3.37** (Multi-valued function of a CRV). In general, if  $y = g(x)$  has  $k$  solutions namely  $y_i = g_i^{-1}(y)$  for  $i = 1, 2, \dots, k$ , then we have

$$p_Y(y) = \sum_{i=1}^k p_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right|.$$

**Theorem 3.10.** Consider a continuous random variable  $X$  with a continuous and monotonically increasing cumulative distribution function  $F_X$ , then  $Y = F_X(X)$  is a uniformly distributed random variable on  $[0, 1]$ .

*Proof.* For  $y \in [0, 1]$ , we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) \\ &= y \implies f_Y(y) = \frac{dF_Y(y)}{dy} = 1. \end{aligned}$$

So, we have

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [0, 1] \\ 0 & \text{if } y \notin [0, 1]. \end{cases}$$

□

**Theorem 3.11.** Consider a continuous random variable  $X$  such that  $X = F_X^{-1}(U)$ , where  $U \sim \mathcal{U}(0, 1)$ , then  $X$  has the probability density function  $f_X(x) = \frac{dF_X(x)}{dx}$ .

*Proof.*

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_U(F_X(x)) \\ &= F_X(x) \implies f_X(x) = \frac{dF_X(x)}{dx}. \end{aligned}$$

□

**Remark 3.5.** In simulating the outcome of a discrete random variable  $X$ , first an outcome of a  $U \sim \mathcal{U}(0, 1)$  random variable is generated and then mapped into a value of  $X$  under the inverse of the CDF. This result is also valid for a continuous random variable so that  $X = F_X^{-1}(U)$  is a random variable with CDF  $F_X(x)$ .

**Example 3.38** (Laplace distribution). Consider a continuous random variable with the following Laplace distribution:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}|x|\right) \text{ for } x \in (-\infty, \infty).$$

Then, its cumulative distribution can be computed as follows:

$$F_X(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}|x|\right) dx.$$

For  $x < 0$ , we have

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\sqrt{\frac{2}{\sigma^2}}x\right) dx = \frac{1}{2} \exp\left(\sqrt{\frac{2}{\sigma^2}}x\right)$$

For  $x \geq 0$ , we have

$$\begin{aligned} F_X(x) &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\sqrt{\frac{2}{\sigma^2}}x\right) dx + \int_0^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}x\right) dx \\ &= 1 - \frac{1}{2} \exp\left(-\sqrt{\frac{2}{\sigma^2}}x\right). \end{aligned}$$

Now, we have  $y = F_X(x)$  for each  $x$ , so

$$y = \begin{cases} \frac{1}{2} \exp\left(\sqrt{\frac{2}{\sigma^2}}x\right) & \text{if } x < 0 \\ 1 - \frac{1}{2} \exp\left(-\sqrt{\frac{2}{\sigma^2}}x\right) & \text{if } x \geq 0. \end{cases}$$

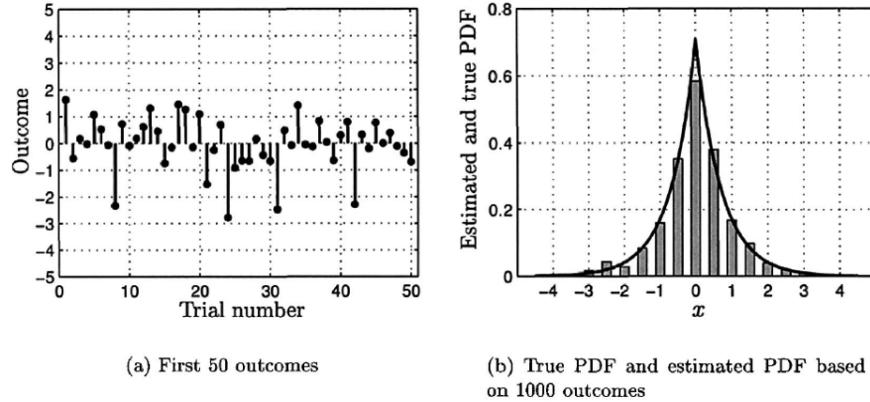
So,  $0 < y < 1/2$  for  $x < 0$  and  $1/2 \leq y < 1$  for  $x \geq 0$ , then we have the inverse map

$$x = \begin{cases} \sqrt{\sigma^2/2} \ln(2y) & \text{for } y \in (0, 1/2) \\ \sqrt{\sigma^2/2} \ln(1 - 2y) & \text{for } y \in [1/2, 1]. \end{cases}$$

Finally to generate the outcome of a Laplacian random variable we can use

$$x = \begin{cases} \sqrt{\sigma^2/2} \ln(2u) & \text{for } u \in (0, 1/2) \\ \sqrt{\sigma^2/2} \ln(1 - 2u) & \text{for } u \in [1/2, 1], \end{cases}$$

where  $u$  is a realization of a uniformly distributed random variable  $\mathcal{U}(0, 1)$ .



**Figure 3.26:** An example of the outcomes of a Laplacian random variable with  $\sigma^2 = 1$ , along with the true PDF (the solid curve) and the estimated PDF (the bar plot).

### 3.3.2 Jointly continuous distribution

In science and in real life, we are often interested in two (or more) random variables at the same time in a random experiment. In such situations the random variables have a joint distribution that allows us to compute probabilities of events involving both variables and understand the relationship between the variables. This is simplest when the variables are independent. When they are not, we use covariance and correlation as measures of the nature of the dependence between them.

**Example 3.39** (Health of a student at IIIT-DWD). Consider an experiment of investigate health of a student at III-TDWD, then we need to measure a list of possible attributes (jointly) such as:

- health of a student  $\rightarrow$  varies randomly in continuum i.e. it may be characterized as a random variable  $H$ .
- weight of a student  $\rightarrow$  varies randomly in continuum i.e. it may be characterized as a random variable  $W$ .
- average blood pressure of a student  $\rightarrow$  varies randomly in continuum i.e. it may be characterized as a random variable  $B$ .

- d. body temperature of a student → varies randomly in continuum i.e. it may be characterized as a random variable  $T$ .
- e. and a few more.

So, using some statistical method, we can come up with a conclusion that

$$\text{health of a student at IIITDWD} = g(H, W, B, T, O).$$

**Definition 3.19** (Jointly continuous random variables). Suppose  $X$  and  $Y$  are two continuous random variables and that  $X$  takes values  $\Omega_X$  and  $Y$  takes values  $\Omega_Y$ . The ordered pair  $(X, Y)$  take values in the product  $\Omega_X \times \Omega_Y$ . The joint probability density function (joint pdf) of  $X$  and  $Y$ ,  $f_{(X,Y)}(x, y)$  is defined as the probability per unit as follows:

$$f_{X,Y}(x, y) = \lim_{(\delta x, \delta y) \rightarrow (0^+, 0^+)} \frac{P(x \leq X \leq x + \delta x, y \leq Y \leq y + \delta y)}{\delta x \delta y}.$$

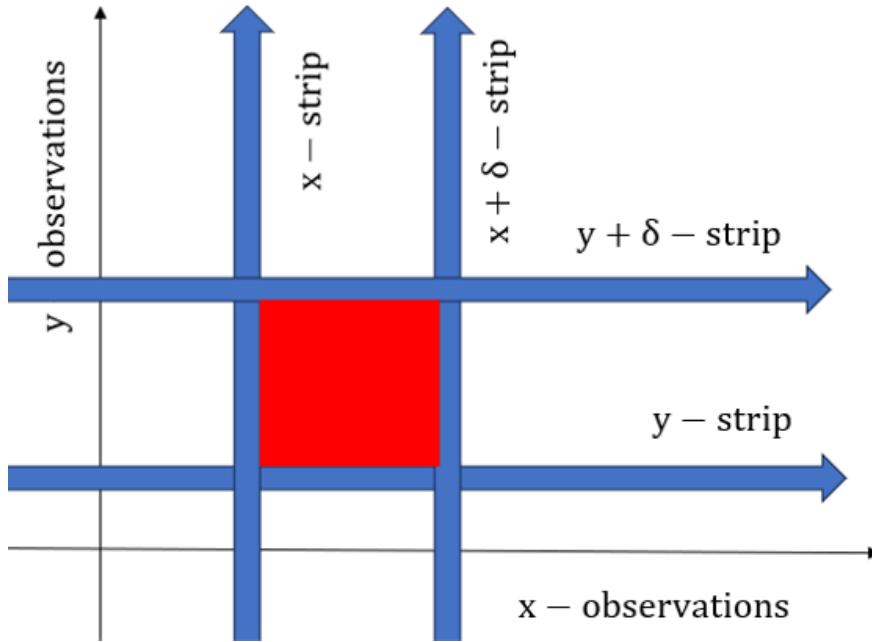
And satisfies the following properties:

1.  $f_{(X,Y)}(x, y) \geq 0$  for any  $(x, y) \in \Omega_X \times \Omega_Y$ ,
2.  $\int_{\{(x,y) \in \Omega_X \times \Omega_Y\}} f_{(X,Y)}(x, y) dx dy = 1$ ,
3.  $P((X, Y) \in B) = \int_{\{(x,y) \in B\}} f_{(X,Y)}(x, y) dx dy$ .

Computationally visualizing of joint PDF over the red region in figure (3.27).

**Definition 3.20** (Conditional PDF for a CRV  $X$  conditioned on  $Y = y$ ). Consider  $X$  and  $Y$  are jointly-distributed and continuous random variables with a statistically dependent, we may want to consider only one of them say  $X$  conditioned on one observation the other, say  $X|Y = y$ . Then, conditional CDF of  $X$ , conditioned on  $Y = y$  is defined as

$$\begin{aligned} f_{X|Y}(x|y) &= \lim_{\delta \rightarrow 0} \frac{P(x \leq X \leq x + \delta | Y = y)}{\delta} = \frac{f_{X,Y}(x, y)}{f_Y(y)} \\ &= \frac{f_{X,Y}(x, y)}{\int_{x \in \mathbb{R}} f_{X,Y}(x, y) dx}. \end{aligned}$$



**Figure 3.27:** A visualizing of computation of joint PDF over the red region.

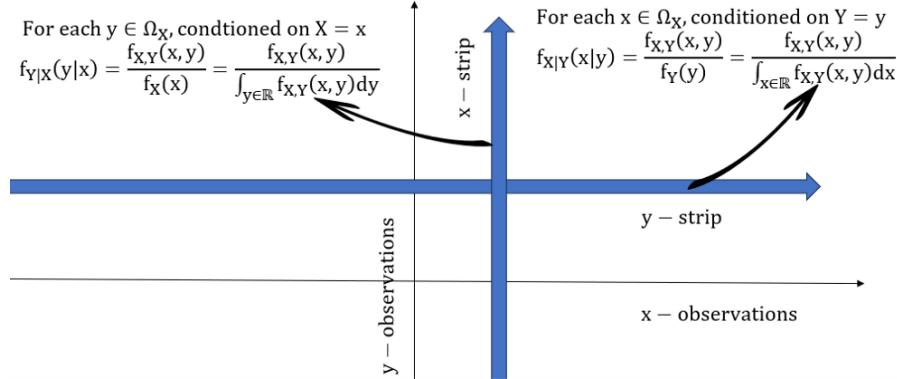
Where derivation of conditional probability  $P(x \leq X \leq x + \delta | Y = y)$  performed as:

$$\begin{aligned}
 P(x \leq X \leq x + \delta | Y = y) &= \lim_{\varepsilon \rightarrow 0} P(x \leq X \leq x + \delta | Y \in (y - \varepsilon, y + \varepsilon)) \\
 &= \lim_{\varepsilon \rightarrow 0} \frac{P(x \leq X \leq x + \delta, y - \varepsilon \leq Y \leq y + \varepsilon)}{P(y - \varepsilon \leq Y \leq y + \varepsilon)} \\
 &\approx \lim_{\varepsilon \rightarrow 0} \frac{f_{X,Y}(x, y)\delta\varepsilon}{f_Y(y)\varepsilon} = \frac{f_{X,Y}(x, y)\delta}{f_Y(y)}.
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 f_{Y|X}(y|x) &= \lim_{\delta \rightarrow 0} \frac{P(y \leq Y \leq y + \delta | X = x)}{\delta} = \frac{f_{X,Y}(x, y)}{f_X(x)} \\
 &= \frac{f_{X,Y}(x, y)}{\int_{y \in \mathbb{R}} f_{X,Y}(x, y) dy}.
 \end{aligned}$$

More explicitly, computational visualization of conditional PDFs in figure (??) as:



**Figure 3.28:** Computationally visualizing conditional PDFs  $f_{Y|X}(y|x)$  &  $f_{X|Y}(x|y)$ .

A geometrical visualization of conditional PDFs including joint and marginal PDFs in figure (3.29) as:

**Example 3.40** (Conditional PDF from standard bivariate PDF). Consider a joint continuously distributed bivariate  $(X, Y)$  with a standard bivariate Gaussian PDF as follows:

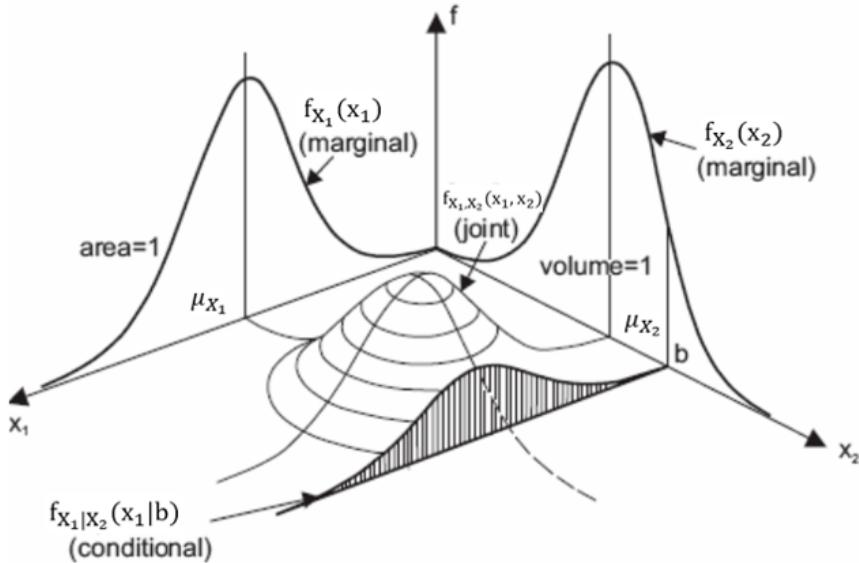
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

and the marginal PDF for X is computed as

$$f_X(x) = \int_{y \in \mathbb{R}} f_{X,Y}(x,y) dy = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Now, the condition PDF of Y conditioned on  $X = x$  can be computed as follows:

$$f_{Y|X}(y|x) = \frac{\frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{Q}{2}\right),$$



**Figure 3.29:** A geometrical visualizing conditional PDF  $f_{X_1|X_2}(x_1|b)$  including joint and marginal PDFs.

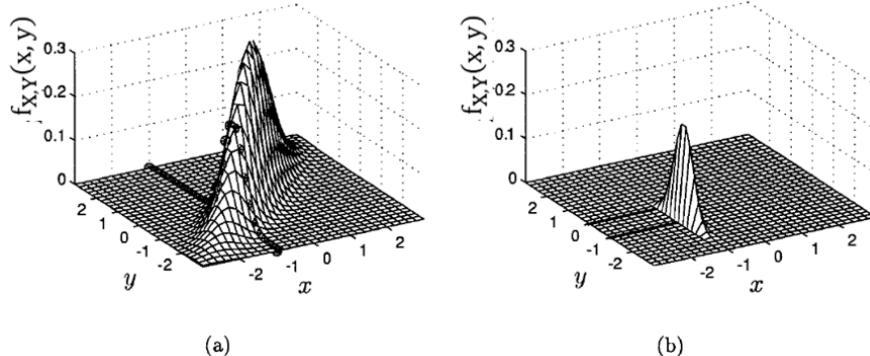
where

$$Q = -\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)} + \frac{x^2}{2} = \frac{(y - \rho x)^2}{(1 - \rho^2)}.$$

So, we can write  $Y|X = x \sim N(x\rho, 1 - \rho^2)$  and visualize in figure (3.30) as follows:

And, visualizing joint marginal and conditional PDFs together in figure (3.31) as follows:

**Example 3.41.** Let S denote the height (in inches) of a randomly chosen father, and T denote the height (in inches) of his son at maturity. Suppose each of S and T has a  $N(\mu, \sigma^2)$  distribution with  $\mu = 69$  and



**Figure 3.30:** Standard bivariate Gaussian PDF with  $\rho = 0.9$  and its cross-section at  $x = -1$ . The normalized cross-section is the conditional PDF.

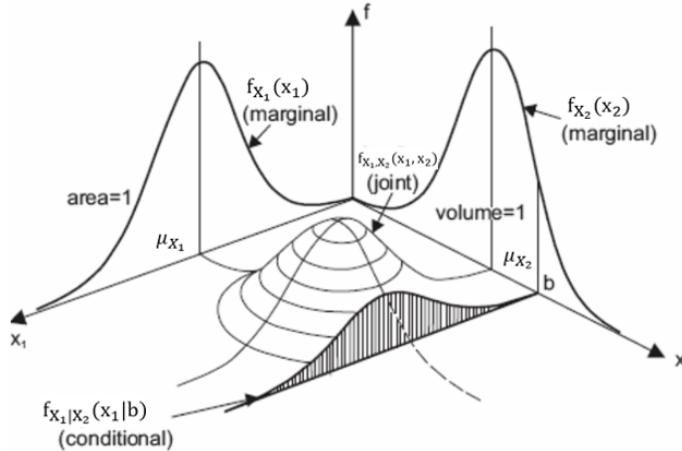
$\sigma = 2$ . Suppose also that the standardized variables  $\frac{S-\mu}{\sigma}$  and  $\frac{T-\mu}{\sigma}$  have a standard bivariate normal distribution with correlation  $\rho = 0.3$ . If Sam has a height of  $S = 74$  inches, the predict the ultimate height  $T$  of his young son Tom.

In standardized units, Sam has height  $X = \frac{S-\mu}{\sigma}$ , which we are given to equal 2.3. Tom's ultimate standardized height is  $Y = \frac{T-\mu}{\sigma}$ . By assumption, before the value of  $X$  was known, the pair  $(X, Y)$  has a standard bivariate normal distribution with correlation  $\rho$ . The conditional distribution of  $Y$  given that  $X = 2.5$  is

$$Y|X = 2.5 \sim N(2.5\rho, 1 - \rho^2).$$

In the original units, the conditional distribution of  $T$  given  $S = 74$  is normal with mean  $\mu + 2.5\rho$  and variance  $(1 - \rho^2)\sigma^2$ , that is, Tom's ultimate height | Sam's height = 74 inches  $\sim N(70.5, 3.64)$ . If I had to make a guess, I would predict that Tom would ultimately reach a height of 70.5 inches.

**Definition 3.21** (Conditional CDF for a CRV  $X$  conditioned on  $Y = y$ ). Consider  $X$  and  $Y$  are jointly-distributed and continuous random variables with a statistically dependent, we may want to consider only one of them say  $X$  conditioned on one observation the other, say  $X|Y = y$ .



**Figure 3.31:** Visualizing joint marginal and conditional PDFs in together.

Then, conditional CDF of X, conditioned on  $Y = y$  is defined as

$$F_{X|Y}(x|y) = \lim_{\varepsilon \rightarrow 0} P(X \leq x | Y \in (y - \varepsilon, y + \varepsilon)) = \int_{-\infty}^x f_{X|Y}(x|y) dx.$$

**Example 3.42.** Consider a joint continuously distributed bivariate  $(X, Y)$  with a standard bivariate Gaussian PDF as follows:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

and the marginal PDF for X is computed as

$$f_X(x) = \int_{y \in \mathbb{R}} f_{X,Y}(x, y) dy = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Now, the condition PDF of Y conditioned on  $X = x$  can be computed as follows:

$$f_{Y|X}(y|x) = \frac{\frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{Q}{2}\right),$$

where

$$Q = -\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} + \frac{x^2}{2} = \frac{(y - \rho x)^2}{(1-\rho^2)}.$$

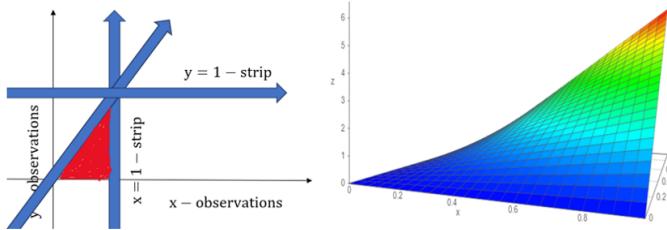
So, we can write  $Y|X = x \sim N(x\rho, 1 - \rho^2)$ . Then

$$F_{Y|X}(y|x) = \int_{-\infty}^y f_{Y|X}(y|x) dy = 1 - \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right).$$

**Example 3.43.** Consider two continuously distributed random variables  $X, Y \sim \mathcal{U}(0, 1)$  and distributed jointly with the following joint PDF:

$$f_{X,Y}(x, y) = \begin{cases} cxy & \text{if } 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then compute (a.)  $c$ , (b.)  $P(Y \leq \frac{X}{2})$  and (c.)  $P(Y \leq \frac{X}{4})$ . And, visualization of domain of joint PDF and surface of the joint PDF in figure (??) as:



**Figure 3.32:** Visualizing domain of joint PDF in left plot and surface of the joint PDF in right plot.

Answer: to compute  $c$ , we have

$$\begin{aligned} 1 &= \int_0^1 \int_0^1 f_{X,Y} dx dy = \int_0^1 \int_y^1 cxy dx dy = \frac{c}{2} \int_0^1 (1 - y^2) y dy = \frac{c}{8} \\ \implies c &= 8. \end{aligned}$$

Now, to compute the probability  $P(Y \leq \frac{X}{2})$ , we have

$$P(Y \leq \frac{X}{2}) = \int_0^1 \int_0^{\frac{x}{2}} 8xy dy dx = \int_0^1 x^3 dx = \frac{1}{4}.$$

Finally, to compute the probability  $P(Y \leq \frac{X}{4})$ , we have

$$P(Y \leq \frac{X}{4}) = \int_0^1 \int_0^{\frac{x}{4}} 8xy dy dx = \frac{1}{4} \int_0^1 x^3 dx = \frac{1}{16}.$$

**Theorem 3.12** (Computation of joint PDF). So, by using the definition of conditional probability for events we have

$$f_{(X,Y)}((x,y)) = f_Y(y)f_{X|Y}(x|y) = f_X(x)f_{Y|X}(y|x).$$

It provides a computational framework to determine joint PMF of the joint discrete random variable  $(X, Y)$ .

**Example 3.44** (Life-time of a spare bulb). Question: a professor uses the overhead projector for his class. The time to failure of a new bulb  $X$  has the exponential PDF  $f_X(x) = \lambda \exp(-\lambda x)u(x)$ , where  $x$  is in hours. A new spare bulb also has a time to failure  $Y$  that is modelled as an exponential PDF. However, the time to failure of the spare bulb depends upon how long the spare bulb sits unused. Assuming the spare bulb is activated as soon as the original bulb fails, the time to activation is given by  $X$ . As a result, the expected time to failure of the spare bulb is decreased as

$$\frac{1}{\lambda_Y} = \frac{1}{\lambda(1 + \alpha x)} \text{ for } \alpha \in (0, 1),$$

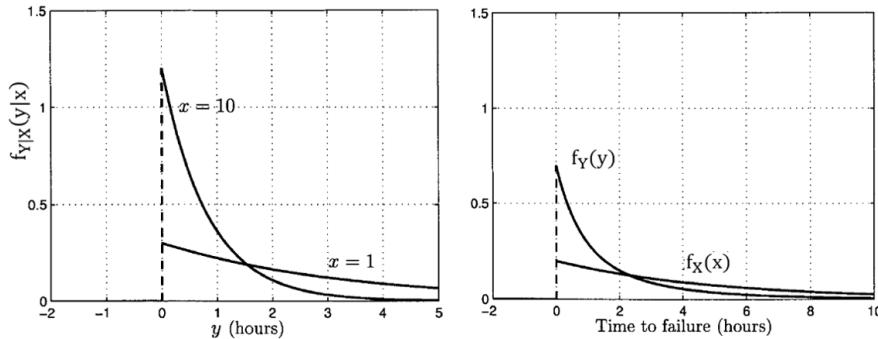
where  $\alpha$  is some factor that indicates the degradation of the unused bulb with storage time. Then compute the PDF of  $Y$ .

Answer: the expected time to failure of the spare bulb decreases as the original bulb is used longer (and hence the spare bulb must sit unused longer). Thus, we model the time to failure of the spare bulb as

$$f_Y(y|x) = \lambda_Y \exp(-\lambda_Y y)u(y) = \lambda(1 + \alpha x) \exp(-\lambda(1 + \alpha x)y) u(y).$$

Then, we compute the desired PDF of  $Y$  as follows:

$$\begin{aligned}
 f_Y(y) &= \int_x f_X(x) f_{Y|X}(y|x) dx = \int_x \lambda^2(1 + \alpha x) e^{-(\lambda(x+(1+\alpha x)y))} dx \\
 &= \lambda^2 \exp(-\lambda y) \int_x (1 + \alpha x) \exp(ax) dx, \quad a = -\lambda(\alpha y + 1) \\
 &= \lambda^2 \exp(-\lambda y) \left( \int_0^\infty \exp(ax) dx + \alpha \int_0^\infty x \exp(ax) \right) \\
 &= \lambda^2 \exp(-\lambda y) \left( -\frac{1}{a} + \frac{\alpha}{a^2} \right), \quad a = -\lambda(\alpha y + 1) \implies \\
 f_Y(y) &= \lambda^2 \exp(-\lambda y) \left( \frac{1}{(\lambda y + 1)} + \frac{\alpha}{(\alpha y + 1)^2} \right) u(y).
 \end{aligned}$$



**Figure 3.33:** Conditional PDF for lifetime of spare bulb with dependence on time to failure  $x$  of original bulb, and PDFs for time to failure of original bulb  $X$  & spare bulb  $Y$ .

**Theorem 3.13** (An another form of total probability). In general to compute probabilities of events it is advantageous to use conditioning arguments whether or not  $X$  and  $Y$  are independent.

$$P(Y \in A) = \int_x P(Y \in A | X = x) f_X(x) dx = \int_x \int_A f_{Y|X}(y|x) dy f_X(x) dx.$$

**Theorem 3.14** (Bayes rule for probability density function). Consider two jointly continuous random variables  $X$  with PDF  $f_X(x)$  and  $Y$  with PDF  $f_Y(y)$  having joint PDF  $f_{X,Y}(x,y)$ , then restatement of the definition of conditional probability density function leads to computation

of a posterior distribution as:

$$f_{(X,Y)}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y) \implies \\ f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{\sum_{\{x \in \Omega_X\}} f_X(x)f_{Y|X}(y|x)},$$

where  $f_X(x) \rightarrow$  prior distribution of  $X$ ,  $f_{Y|X}(y|x) = f(x) \rightarrow$  is the data generating process as a likelihood of  $Y = y$  conditioned on  $X = x$  and  $f_Y(y) = \sum_{\{x \in \Omega_X\}} f_X(x)f_{Y|X}(y|x) \rightarrow$  probability of observing  $Y = y$  or evidence.

**Definition 3.22** (Independent jointly continuous random variables). Jointly-distributed continuous random variables  $X$  and  $Y$  are independent if observation of one random variable before the other is not going to affect their probability density functions or their joint PDF is the product of the marginal PDF's:

$$f_{X|Y}(x|y) = f_X(x) \text{ or } f_{Y|X}(y|x) = f_Y(y),$$

or

$$f_{(X,Y)}(x,y) = f_X(x)f_Y(y) \text{ for any } (x,y) \in \Omega_X \times \Omega_Y.$$

### 3.3.3 Derived distribution of a continuous random variable

**Definition 3.23** (Derived distribution of a continuous random variable). Consider a continuous random variable  $X$  with its PDF  $f_X(x)$  and a transformation  $Y = g(X)$  as a function of  $X$  such that  $\Omega_Y$  is continuum in nature. Then probability density function of  $Y$  can be computed as follows:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \\ \implies f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dg^{-1}(y)}{dy} \right|.$$

So, computation of derived distribution  $p_Y(y)$  of a continuous random variable  $X$  (with  $\Omega_Y$  as continuum in nature) follows two steps:

1. Determining all possible observed value of the derived random variable  $Y$  i.e.  $\Omega_Y$ ,

2. Computing the derived distribution as follows:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dg^{-1}(y)}{dy} \right|.$$

**Example 3.45** (Derived distribution of a CRV  $X \sim Unif(0,1)$ ). If  $X \sim Unif(0,1)$  and  $Y = \sqrt{X}$ , then derived PDF  $f_Y(y)$  can be computed as:

1. As  $X \sim Unif(0,1)$  and  $Y = \sqrt{X}$ , so  $\Omega_X = [0,1]$  transformed to  $\Omega_Y = [0,1]$  and for each  $x \in \Omega_X, y = \sqrt{x} \implies x = g^{-1}(y) = y^2$ .
2. Now, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\sqrt{X} \leq y) = P(X \leq y^2) = \int_0^{y^2} dx = y^2, \\ \implies f_y(y) &= 2y \text{ for } y \in [0,1]. \end{aligned}$$

Alternatively

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) = P(X \leq y^2) = F_X(y^2), \implies \\ f_y(y) &= \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dy^2}{dy} \right| = 1 \times 2y = 2y. \end{aligned}$$

**Example 3.46.** Question: if  $X \sim Unif(-1,1)$  and  $Y = |X|$ , then compute the derived distribution  $f_Y(y)$  of  $Y = g(X) = |X|$ .

Answer: derived distribution  $f_Y(y)$  of  $Y = g(X) = |X|$  can be computed as:

1. As  $X \sim Unif(-1,1)$  and  $Y = |X|$ , so  $\Omega_X = [-1,1]$  transformed to  $\Omega_Y = [0,1]$  and for each  $x \in \Omega_X, y = |x| \implies x = g^{-1}(y) = \{-y, y\}$ .
2. Now, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(|X| \leq y) = P(X \leq y \in \{-y, y\}) \\ &= P(-y \leq X \leq y) = P(X \leq y) - P(X \leq -y), \\ \implies f_y(y) &= 1 \text{ for } y \in [0,1]. \end{aligned}$$

Alternatively

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(|X| \leq y) = P(X \leq y \in \{-y, y\}) \\ &= P(-y \leq X \leq y) = F_X(y) - F_X(-y), \\ \implies f_y(y) &= \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dg^{-1}(y)}{dy} \right| = \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1. \end{aligned}$$

**Example 3.47.** Question: If  $X \sim \text{Unif}(-1, 1)$  and  $Y = X^2$ , then compute the derived distribution  $f_Y(y)$  of  $Y = g(X) = X^2$ .

Answer: derived distribution  $f_Y(y)$  of  $Y = g(X) = X^2$  can be computed as:

1. As  $X \sim \text{Unif}(-1, 1)$  and  $Y = X^2$ , so  $\Omega_X = [-1, 1]$  transformed to  $\Omega_Y = [0, 1]$  and for each  $x \in \Omega_X, y = x^2 \implies x = g^{-1}(y) = \{-\sqrt{y}, \sqrt{y}\}$ .
2. Now, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(X \leq y \in \{-\sqrt{y}, \sqrt{y}\}) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}), \\ \implies f_y(y) &= \frac{1}{4\sqrt{y}} + \frac{1}{4\sqrt{y}}. \end{aligned}$$

Alternatively

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(X \leq y \in \{-\sqrt{y}, \sqrt{y}\}) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ \implies f_y(y) &= \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dg^{-1}(y)}{dy} \right| = \frac{1}{2} \left( \frac{1}{2\sqrt{y}} + \frac{1}{2\sqrt{y}} \right) = \frac{1}{2\sqrt{y}}. \end{aligned}$$

**Example 3.48.** Question: if  $X \sim N(0, 1)$  and  $Y = g(X) = X^2$ , then compute the derived distribution  $f_Y(y)$  of  $Y = g(X) = X^2$ .

Answer: derived distribution  $f_Y(y)$  of  $Y = g(X) = X^2$  can be computed as:

1. As  $X \sim N(0, 1)$  and  $Y = g(X) = X^2$ , so  $\Omega_X = (-\infty, \infty)$  transformed to  $\Omega_Y = [0, \infty)$ .

2. Now, for  $y \in [0, \infty)$ , we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(X \leq y \in \{-\sqrt{y}, \sqrt{y}\}) \\ &= P(X \leq -\sqrt{y}) + P(X \leq \sqrt{y}) \\ \implies f_y(y) &= f_X(x) \left( \frac{1}{2\sqrt{y}} + \frac{1}{2\sqrt{y}} \right) = \frac{1}{\sqrt{2\pi y}} \exp(-y/2). \end{aligned}$$

Alternatively

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) = P(X \leq y \in \{-y, y\}) = F_X(-y) + F_Y(y), \\ \implies f_y(y) &= \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \left| \frac{dy^2}{dy} \right| = f_X(x) \left( \frac{1}{2\sqrt{y}} + \frac{1}{2\sqrt{y}} \right) \\ &= \frac{1}{\sqrt{2\pi y}} \exp(-y/2). \end{aligned}$$

And,  $Y$  does not observe any value in  $(-\infty, 0)$ , so, we have

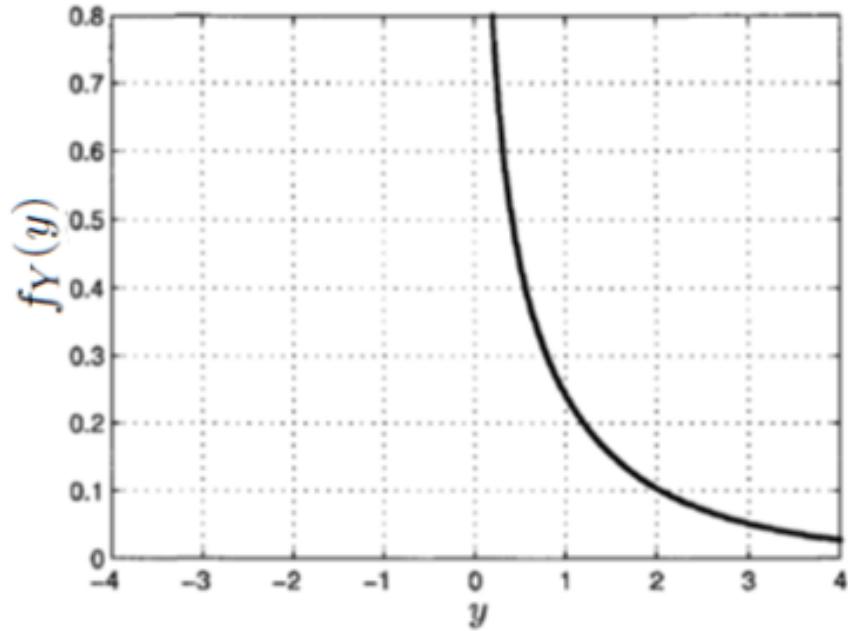
$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} \exp(-y/2) & \text{if } y \geq 0 \\ 0, & \text{if } y < 0. \end{cases}$$

**Example 3.49.** If  $X \sim N(0, 1)$  and  $Y = g(X) = \exp(X)$ , then compute the derived distribution  $f_Y(y)$  of  $Y = g(X) = \exp(X)$ .

Answer: derived distribution  $f_Y(y)$  of  $Y = g(X) = \exp(X)$  can be computed as:

1. As  $X \sim N(0, 1)$  and  $Y = g(X) = \exp(X)$ , so  $\Omega_X = (-\infty, \infty)$  transformed to  $\Omega_Y = (0, \infty)$ .
2. Now, for  $y \in (0, \infty)$ , we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\exp(X) \leq y) = P(X \leq \ln(y)) = F_X(\ln(y)) \\ \implies f_y(y) &= \frac{dF_X(\ln(y))}{d\ln(y)} \left| \frac{d\ln(y)}{dy} \right| = f_X(\ln(y)) \frac{1}{y} \\ &= \frac{1}{\sqrt{2\pi y}} \exp(-(\ln(y))^2/2). \end{aligned}$$



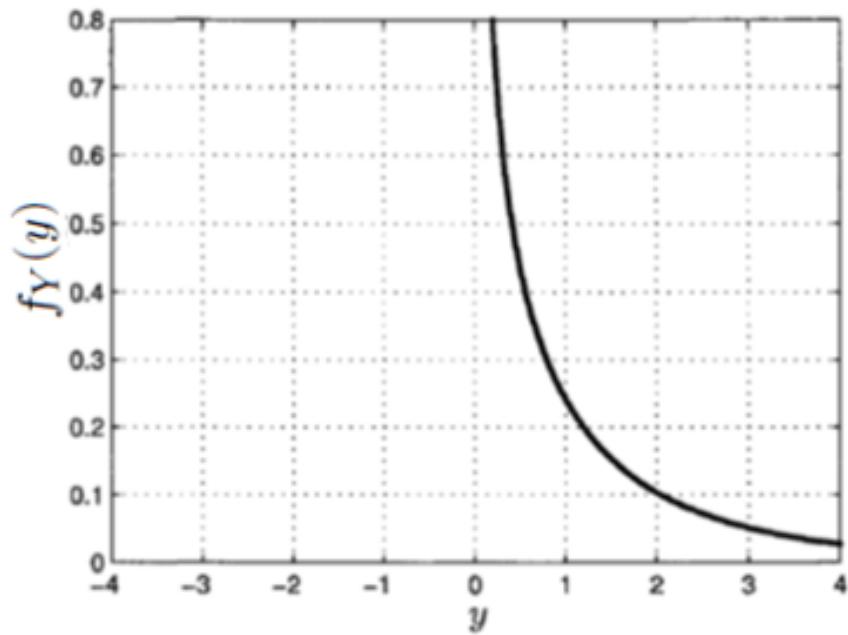
**Figure 3.34:** Probability density function for  $Y = X^2$ ,  $X \sim N(0, 1)$ .

Alternatively

$$\begin{aligned}
 F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) \\
 &= F_X(g^{-1}(y)) = P(X \leq y \in \{-y, y\}) = F_X(-y) + F_Y(y), \\
 \implies f_y(y) &= \frac{dF_X(\ln(y))}{d\ln(y)} \left| \frac{d\ln(y)}{dy} \right| = f_X(\ln(y)) \frac{1}{y} \\
 &= \frac{1}{\sqrt{2\pi}y} \exp\left(-\frac{(\ln(y))^2}{2}\right).
 \end{aligned}$$

And,  $Y$  does not observe any value in  $(-\infty, 0)$ , so, we have

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}y} \exp\left(-\frac{(\ln(y))^2}{2}\right) & \text{if } y > 0 \\ 0 & \text{if } y \leq 0. \end{cases}$$



**Figure 3.35:** Probability density function for  $Y = X^2$ ,  $X \sim N(0, 1)$ .

This PDF is called the log-normal PDF. It is frequently used as a model for a quantity that is measured in decibels (dB) and which has a normal PDF in dB quantities.

**Corollary 3.15** (PDF of a function of jointly continuous random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly continuous and independent random variables and a function  $Z = \frac{Y}{X}$ , then PDF is  $Z$  is computed as follows:

$$f_Z(z) = \int_x f_X(x) f_Y(xz) |x| dx.$$

*Proof.* Derivation as follows:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(Y \leq zx|X=x) f_X(x) dx \\ \implies f_Z(z) &= \frac{\partial F_Z(z)}{\partial z} = \int_x \frac{\partial F_{Y|X}(zx|X=x)}{\partial z} f_X(x) dx \\ &= \int_x f_{Y|X}(zx|X=x)|x|f_X(x)dx = \int_x f_X(x)f_Y(zx)|x|dx. \end{aligned}$$

□

**Corollary 3.16** (PDF of a function of jointly continuous random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly continuous and independent random variables and a function  $Z = XY$ , then PDF is  $Z$  is computed as follows:

$$f_Z(z) = \int_x f_X(x)f_Y(z/x) \frac{1}{|x|} dx.$$

*Proof.* Derivation as follows:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P\left(Y \leq \frac{z}{X}\right) = \int_x P\left(Y \leq \frac{z}{x}|X=x\right) f_X(x) dx \\ \implies f_Z(z) &= \frac{\partial F_Z(z)}{\partial z} = \int_x \frac{\partial F_{Y|X}(z/x|X=x)}{\partial z} f_X(x) dx \\ &= \int_x f_{Y|X}(z/x|X=x) \frac{1}{|x|} f_X(x) dx = \int_x f_X(x)f_Y(z/x) \frac{1}{|x|} dx. \end{aligned}$$

□

**Theorem 3.17** (PDF of a sum of jointly continuous random variables  $(X, Y)$ ). Consider  $X$  and  $Y$  are two jointly continuous and independent random variables and a function  $Z = X + Y$ , then PDF of  $Z$  is computed as follows:

$$f_Z(z) = \int_x f_X(x)f_Y(z-x) dx.$$

*Proof.* Derivation as follows:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X + Y \leq z) = \int_x P(Y \leq z - x|X=x) f_X(x) dx \implies \\ f_Z(z) &= \frac{\partial F_Z(z)}{\partial z} = \int_x f_X(x) \frac{\partial F_{Y|X}(z-x|X=x)}{\partial z} dx = \int_x f_X(x)f_{Y|X}(z-x|X=x) dx \\ &= \int_x f_X(x)f_Y(z-x) dx. \end{aligned}$$

□

**Example 3.50.** Suppose X and Y are uniformly distributed on  $[0, 1]$  and independent random variables. Find the PDF of  $Z = X + Y$ .

Solution: We have  $X, Y \sim \mathcal{U}(0, 1)$ , then

$$f_X(x) = \begin{cases} 1 & x \in [0, 1], \\ 0 & x \notin [0, 1]. \end{cases}$$

and

$$f_Y(y) = \begin{cases} 1 & y \in [0, 1], \\ 0 & y \notin [0, 1]. \end{cases}$$

Now,  $Z = X + Y$ , so  $f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$  can be computed case-wise as follows:

Case-I: If  $0 \leq z \leq 1$ , then  $0 \leq x \leq z$ , so we have

$$f_Z(z) = \int_0^z f_X(x)f_Y(z-x)dx = \int_0^z 1 \cdot 1 \cdot dx = z$$

Case-II: If  $1 < z \leq 2 \equiv 0 < z-1 \leq 1$  then  $z-1 \leq x \leq 1$ , so, we have

$$f_Z(z) = \int_{z-1}^1 f_X(x)f_Y(z-x)dx = \int_{z-1}^1 1 \cdot 1 \cdot dx = 2 - z$$

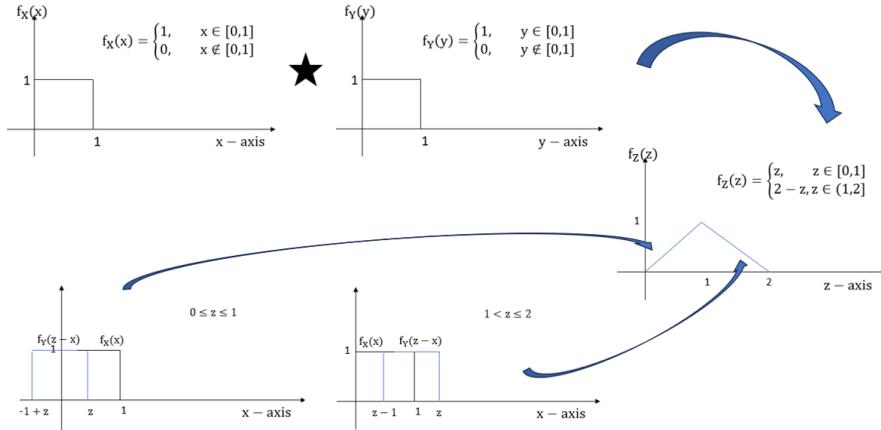
**Example 3.51.** Backdrop: consider two jointly continuous and independent random variables X and Y uniformly distributed in  $[0, 1]$ .

Question: Let  $Z = XY$ , then compute CDF and PDF of Z.

Answer: we have  $X \sim \mathcal{U}(0, 1)$  and  $Y \sim \mathcal{U}(0, 1)$  and independent to each other. So, we get

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{if } x \notin [0, 1]. \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} 1 & \text{if } y \in [0, 1] \\ 0 & \text{if } y \notin [0, 1]. \end{cases} \implies$$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = \begin{cases} 1 & \text{if } (x, y) \in [0, 1] \times [0, 1] \\ 0 & \text{if } (x, y) \notin [0, 1] \times [0, 1]. \end{cases}$$



**Figure 3.36:** Convolution of two uniform and independent distributions on \$[0, 1]\$.

Now, cumulative distribution function of \$Z\$ is given by

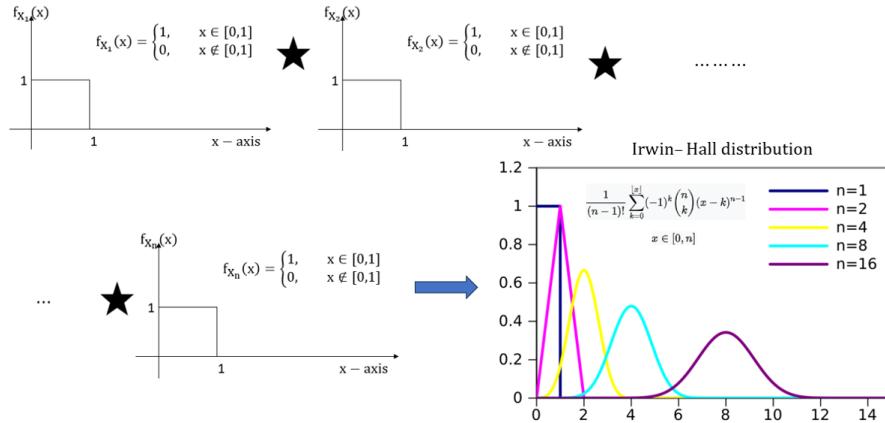
$$\begin{aligned}
 F_Z(z) &= P(Z \leq z) = P(XY \leq z) = P\left(X \leq \frac{z}{Y}\right) = \int_0^1 \int_0^{\min(1, \frac{z}{y})} dx dy \\
 &= \int_0^1 \int_0^{\min(1, \frac{z}{y})} dx dy = \int_0^1 \min\left(1, \frac{z}{y}\right) dy \\
 &= \int_0^z dy + \int_z^1 \frac{z}{y} dy = z - z \ln(z) \implies f_Z(z) = -\ln(z) = \ln\left(\frac{1}{z}\right).
 \end{aligned}$$

Alternatively, using conditioning and law of total probability, we get

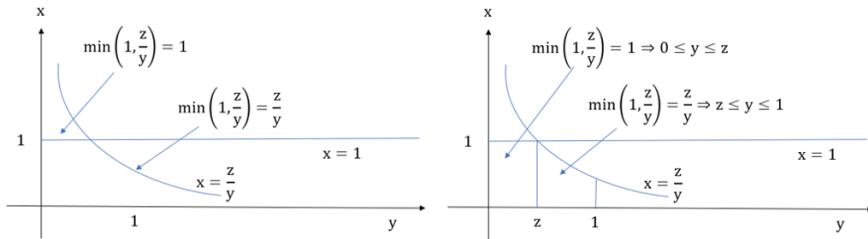
$$\begin{aligned}
 F_Z(z) &= P(Z \leq z) = P(XY \leq z) = P\left(X \leq \frac{z}{Y}\right) \\
 &= \int_0^1 P\left(X \leq \frac{z}{y} | Y = y\right) f_Y(y) dy = \int_0^1 \min\left(1, \frac{z}{y}\right) dy \\
 &= \int_0^z dy + \int_z^1 \frac{z}{y} dy = z - z \ln(z) \implies f_Z(z) = -\ln(z) = \ln\left(\frac{1}{z}\right).
 \end{aligned}$$

### 3.3.4 Expectation and variance of continuous random variables

**Definition 3.24** (Expectation of a continuous random variable). Consider a continuous random variable with a continuous distribution or



**Figure 3.37:** Convolution of  $n$  uniform and independent distributions on  $[0, 1]$ .



**Figure 3.38:** In left plot there are depiction of limits for inner integrating element  $dx$  and in the right plot there are depiction of limits for outer integrating element  $dy$ .

probability density function  $f_X(x)$ . The expectation,  $E(X)$  is defined as:

$$E(X) = \int_{\{x \in \Omega_X\}} x f_X(x) dx.$$

**Definition 3.25 (Expected value rule).** Consider a continuous random variable  $X$  with PDF  $f_X(x)$  and let  $Y = g(X)$  be a function of  $X$ . The

expectation,  $E(g(X))$  is defined as:

$$\begin{aligned} E(g(X)) &= \int_{\{y \in \Omega_Y\}} y f_Y(y) dy = \int_{\{y \in \Omega_Y\}} y \left( \int_{g^{-1}(y) = \{x | g(x) = y\}} f_X(x) dx \right) dy \\ &= \int_{\{x \in \Omega_X\}} g(x) f_X(x) dx. \end{aligned}$$

The expectation tells us important information about the average value of a random variable, but there is a lot of information that it doesn't provide. If you are investing in the stock market, the expected rate of return of some stock is not the only quantity you will be interested in, you would also like to get some idea of the riskiness of the investment. The typical size of the fluctuations of a random variable around its expectation is described by another summary statistic, the variance.

**Definition 3.26** (Variance of a continuous random variable). For a continuous random variable  $X$  with a PDF  $f_X(x)$ , the variance,  $Var(X)$  is defined as:

$$Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2.$$

provided that this quantity exists. We can also treat variance,  $Var(X)$ , of  $X$  as an expected value  $E(g(X))$  of the function  $g(X) = (X - E(X))^2$ . That is variance is a measure of how much the distribution of  $X$  is spread out about its mean and only randomness gives rise to variance. Statisticians often prefer to use the standard deviation (in practice for interpretation and visualization) rather than variance (in principle for theoretical analysis and understanding) as a measure of spread.

**Definition 3.27** (Covariance of two jointly continuous random variables). Consider a jointly continuous random variable  $(X, Y)$  with joint PDF  $f_{X,Y}(x, y)$ , where  $X$  and  $Y$  are associated with the same random experiment, then covariance of  $(X, Y)$  is defined as:

$$\begin{aligned} Cov(X, Y) &= E((X - E(X))(Y - E(Y))) = E(XY - XE(Y) - YE(X) \\ &\quad + E(X)E(Y)) = E(XY) - E(X)E(Y). \end{aligned}$$

Furthermore, the variance-covariance matrix is given by

$$C_{X,Y} = \begin{pmatrix} Var(X) & Cov(X, Y) \\ Cov(X, Y) & Var(Y) \end{pmatrix}.$$

**Theorem 3.18.** If X and Y are independent and jointly continuous random variables whose expectations exist, then

$$a. E(XY) = E(X)E(Y) \text{ and } b. Cov(X, Y) = 0.$$

**Definition 3.28** (Conditional expectation). In a probabilistic model if a certain event A has already occurred, then we define conditional expectation of the possible observations of a continuous random variable X conditioned on the event A as follows:

$$E(X|A) = \int_{\{x \in \Omega_X\}} xf_{X|A}(x|A)dx.$$

Moreover, if the event A is characterized by another continuous random variable Y in the same experiment as  $Y = y$ , then the conditional expectation of X conditioned on  $Y = y$  is defined as follows:

$$E(X|Y = y) = \int_{\{x \in \Omega_X\}} xf_{X|Y}(x|y)dx = g(y).$$

That is for each possible observation  $Y = y$ ,  $E(X|Y = y) = g(y)$ , and hence conditional expectation defines a random variable (as a function of random variable Y) as:

$$g(Y) = E(X|Y).$$

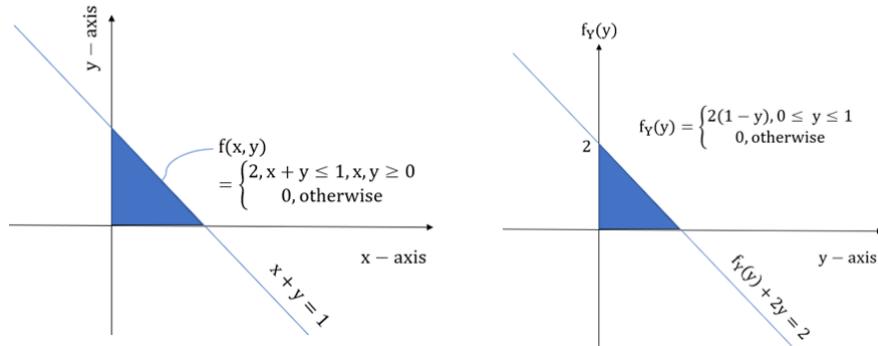
**Example 3.52.** Question: consider two jointly continuous random variables  $(X, Y)$  with a joint PDF  $f_{X,Y} = 2$  over a triangle region  $X + Y \leq 1, X \geq 0, Y \geq 0$ . Compute the conditional expectation  $E(X|Y = y)$  conditioned on the observation  $Y = y$ .

Answer: we have the joint PDF as

$$f_{X,Y}(x, y) = \begin{cases} 2 & x + y \leq 1, x \geq 0, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Now, we compute marginal PDF  $f_Y(y)$  (for  $0 \leq y \leq 1$ ) from the given joint PDF  $f_{X,Y}(x, y)$  as:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \int_0^{1-y} 2dx = 2(1 - y).$$



**Figure 3.39:** In left plot there is a joint region of the joint  $f_{X,Y} = 2$ , in right plot its marginal PDF  $f_Y(y)$ .

And the condition PDF  $f_{X|Y}(x|y)$ , conditioned on  $Y = y$ , can be computed as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2}{2(1-y)} = \frac{1}{1-y}, \quad 0 \leq y \leq 1.$$

Then, we compute the conditional expectation  $E(X|Y = y)$ , conditioned on  $Y = y$ , can be computed as:

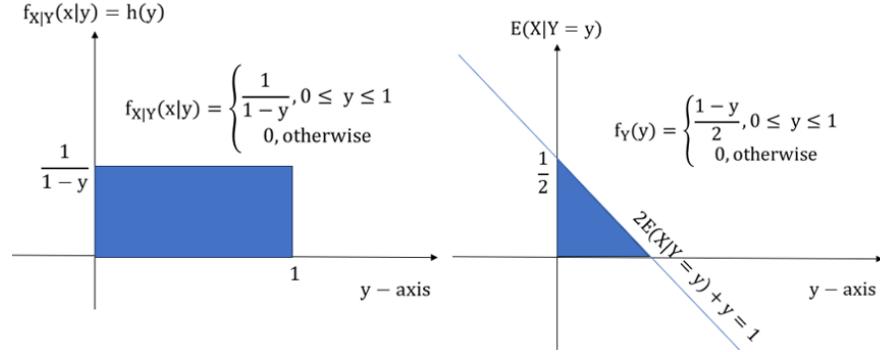
$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx = \int_0^{1-y} x \frac{1}{(1-y)} dx = \frac{1-y}{2}.$$

As  $y$  varies, so does the conditional expectation  $E(X|Y = y)$ , and hence, we have the conditional expectation as a random variable as:

$$\begin{aligned} E(X|Y) &= \frac{1-Y}{2} \implies E(X) = \frac{1}{2}(1 - E(Y)), \quad E(X) = E(Y) \implies \\ E(X) &= \frac{1}{3}. \end{aligned}$$

**Theorem 3.19** (Law of iterated expectation). If  $\{B_1, B_2, \dots\}$  be a partition of  $\Omega$  such that  $P(B_i) > 0 \forall i$ , and  $X$  is a continuous random variable in the same experiment, then

$$E(X) = \sum_{i \geq 1} E(X|B_i)P(B_i).$$



**Figure 3.40:** In left plot there is a conditional PDF  $f_{X|Y}(x|y)$ , in right plot its conditional expectation  $E(X|Y = y)$ .

Moreover, if  $Y$  is a continuous random variable such that  $Y^{-1}(y_i) = B_i = \{\omega | Y(\omega) = y_i\}$ , then

$$E(X) = \int_{\{y \in \Omega_Y\}} E(X|Y = y) f_Y(y) dy = E(E(X|Y)).$$

**Theorem 3.20.** Consider two jointly continuous random variables  $(X, Y)$ . Then, the variance of conditional expectation and expected value of conditional variance are related as:

$$\text{Var}(X) = \text{Var}(E(X|Y)) + E(\text{Var}(X|Y)).$$

*Proof.* We have

$$\begin{aligned} \text{Var}(X) &= E((X - E(X))^2) = E((X - E(X|Y) + E(X|Y) - E(X))^2) \\ &= E((X - E(X|Y))^2) + 2E((X - E(X|Y))(E(X|Y) - E(X))) \\ &\quad + E((E(X|Y) - E(X))^2) \end{aligned}$$

□

**Example 3.53** ([Bertsekas and Tsitsiklis, 2008]). Question: we start with a stick of length  $\ell$ . We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the stick that we were left with. Compute the expected length of the stick that we are left with, after breaking twice.

Answer: Let  $Y$  be the length of the stick after we break for the first time. Let  $X$  be the length after the second time. We have  $E(X|Y) = \frac{Y}{2}$ , since the breakpoint is chosen uniformly over the length  $Y$  of the remaining stick. For a similar reason, we also have  $E(Y) = \frac{\ell}{2}$ . Then, the expected value of  $X$  can be computed as:

$$E(X) = E(E(X|Y)) = E(Y/2) = \frac{\ell}{4}.$$

### 3.4 General random variable

It may be of interest to note that discrete and continuous random variables can be subsumed under the topic of a general random variable. There exists the mathematical machinery to analyse both types of random variables simultaneously. This theory is called measure theory.

# 4

---

## Large sample theory

---

Till now, we have described and computed the exact probability of events based on known probability mass functions (PMFs)/probability density functions (PDFs) and the implementation of its summation/integration. Also of importance were the methods to determine the moments/statistic/parameters of these random variables. In many practical situations the PMF/PDF may be unknown or the summation/integration may not be easily carried out. It would be of great utility, therefore, to be able to approximate the desired quantities using much simpler methods e.g. random sampling and point estimation. Law of large numbers asserts that the sample mean (the average of IID copies) converges to the expected value, a number, of each random variable in the average. Also, it provides a justification for the relative frequency interpretation of probability. Central limit theorem asserts that a properly normalized sum of I.I.D. random variables converges to a Gaussian random variable. Alternatively, it can be demonstrated that the repeated convolution of PDFs produces a Gaussian PDF. So, in this section, we discuss large sample theory of sample statistic, when distribution of a measurement (or a random variable) is unknown. It discusses one (i.e. large sample size) of two important concepts

to understand, obverse and utilize Big-Data. Another one (i.e. high-dimension) will be discussed in the next chapter (5).

## 4.1 Sample statistic

Consider a random variable (or a measurement)  $X$  with unknown distribution, then it is very difficult to compute statistic or parameter of the observed values under the unknown distribution. Now, we need to define sample statistic instead with respect to a random sample of size  $n$  (i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  independent and identically distributed (I.I.D.) copies of  $X$ ).

**Definition 4.1** (Sample mean). Consider a random variable (or a measurement)  $X$  with unknown distribution. Assume a random sample of size  $n$  (i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  independent and identically distributed (I.I.D.) copies of  $X$ ), then sample mean is defined as follows:

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

The sample mean is an estimator of the mean  $\mu$  of the measurement  $X$  and its bias is defined as follows:

$$Bias(\bar{X}) = E(\bar{X}) - \mu.$$

Consider a realization/observation of the random sample of size  $n$ ,  $(x_1, x_2, \dots, x_n)$ , then the corresponding realization of the sample mean as  $\bar{x}$  returns the sample average

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

**Definition 4.2** (Sample variance). Consider a random variable (or a measurement)  $X$  with unknown distribution. Assume a random sample of size  $n$  (i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  independent and identically distributed (I.I.D.) copies of  $X$ ), then sample variance is defined as follows:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right).$$

The sample variance is an estimator of the variance  $\sigma^2$  of the measurement  $X$  and its bias is defined as follows:

$$\text{Bias}(S^2) = E(S^2) - \sigma^2.$$

Consider a realization/observation of the random sample of size  $n, (x_1, x_2, \dots, x_n)$ , then the corresponding realization of the sample variance as  $S^2$  returns the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

Alternatively, we can have another definition of sample variance as follows:

$$\begin{aligned} S^2 &= \hat{\text{Var}}(X) = \bar{X}^2 - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \frac{1}{n} \sum_{i=1}^n (X_i^2 - \bar{X}^2). \end{aligned}$$

**Definition 4.3** (Sample standard deviation). In the above definition (4.2),  $S^2$  is a sample variance of  $X$ , so the sample standard deviation is defined as

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

Further, we have

$$0 < \text{Var}(S^2) = E(S^2) - (E(S))^2 = \sigma^2 - (E(S))^2 \implies E(S) < \sigma.$$

So, the sample standard deviation  $S$  is a biased estimator of the standard deviation  $\sigma$  with a bias

$$\text{Bias}(S) = E(S) - \sigma < 0.$$

Consider a realization/observation of the random sample of size  $n, (x_1, x_2, \dots, x_n)$ , then the corresponding realization of the sample standard deviation as  $S$  returns the sample standard error

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

**Example 4.1** (Average mark in the course MA201). Consider  $X_1, X_2, X_3, X_4, X_5$  are the marks in the different evaluation components (e.g. quiz-1,quiz-2, assignment, mid-term, end-term respectively) i.e. we have a random sample of size 5 (upto normalization). Then, the sample mean  $\bar{X} = \frac{X_1+X_2+\dots+X_5}{5}$  with 70 realization in section-A and 71 realization in section-B and  $\bar{X}$  is a unbiased estimator of the average mark in the course MA201.

**Example 4.2.** Backdrop: let  $T$  be the time that is needed for a specific task in a factory to be completed. In order to estimate the mean and variance of  $T$ , we observe a random sample  $(T_1, T_2, \dots, T_6)$ . Thus,  $T_i$ 's are i.i.d. and have the same distribution as  $T$ .

Question: we obtain the following values (in minutes) as a realization/observation of the random sample of size 6:

$$(t_1, t_2, t_3, t_4, t_5, t_6) = (18, 21, 17, 16, 24, 20).$$

Then compute the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

Answer: as we have

$$\bar{T} = \frac{T_1 + T_2 + \dots + T_6}{6} \quad S^2 = \frac{1}{5} \left( \sum_{i=1}^n T_i^2 - 6\bar{T}^2 \right)$$

So, with respect to the realization/observation  $(t_1, t_2, t_3, t_4, t_5, t_6) = (18, 21, 17, 16, 24, 20)$  of the random sample of size 6, we have

$$\bar{t} = \frac{t_1 + t_2 + \dots + t_6}{6} \quad S^2 = \frac{1}{5} \left( \sum_{i=1}^n t_i^2 - 6\bar{t}^2 \right)$$

So, we have the following computations:

$$\bar{T} = \bar{t} = 19.33, \quad S^2 = 8.67, \quad S = \sqrt{S^2} = 2.94.$$

#### 4.1.1 Methods of estimation for sample statistic

Various methods of estimation for sample statistic such as:

- a. maximum likelihood estimation,
- b. method of moments,

- c. max a-posteriori estimation,
- d. least squared estimation

will be discussed in detail in next chapter (5).

#### **4.1.2 Method of sampling**

In practice (e.g. in everyday activity or scientific research), our understanding and applicability are based on observation of some relevant samples. For example, a person (with several years of experience with an institute) can often determine an opinion, of an institute that conducts thousands of transactions everyday, by one or two encounters. In other scenario, a traveller who spends 10 days in a country and then proceeds to write a book telling the inhabitants how to revive their industries, reform their political system, balance their budget, and improve the food in their hotels is a familiar figure of fun. But in reality he differs from the political scientist who devotes 20 years to living and studying in the country only in that he bases his conclusions on a much smaller sample of experience and is less likely to be aware of the extent of his ignorance. In every branch of science we lack the resources to "study more than a fragment of the phenomena that might advance our knowledge. A list of the principal advantages of sampling as compared with complete enumeration.

- a. Reduced cost: if data are secured from only a small fraction of the aggregate, expenditures may be expected to be smaller than if a complete census is attempted.
- b. Greater speed: for the same reason, the data can be collected and summarized more quickly with a sample than with a complete count. This may be a vital consideration when the information is urgently needed.
- c. Greater scope: in certain types of inquiry, highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete census may then be impracticable: the choice lies between obtaining the information by sampling or

not at all. Thus surveys which rely on sampling have more scope and flexibility as to the types of information that can be obtained. On the other hand, if information is wanted for many subdivisions or segments of the population, it may be found that a complete enumeration offers the best solution.

- d. Greater accuracy: because personnel of higher quality can be employed and can be given intensive training, a sample may actually produce more accurate results than the kind of complete enumeration that it is feasible to take.

There are various methods of sampling:

- a. Random sampling
- b. Other sampling
- c. Issues with sampling such as sampling bias.

## 4.2 Probabilistic inequalities

There are three fundamental probabilistic inequalities:

- a. Markov inequality
- b. Chebushev inequality
- c. Chernoff inequality

### 4.2.1 Probabilistic inequalities and non-asymptotic sample theory

The mean and variance of a random variable indicate the average value and variability of the outcomes of a repeated experiment. As such, they summarize important information about the distribution (i.e. PMF/PDF). However, they are not sufficient to determine probabilities of events. For example, consider the PDFs:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad f_Y(y) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|y|\right).$$

both have  $E(X) = 0 = E(Y)$  (due to symmetry about origin) and  $\text{var}(X) = 1 = \text{Var}(Y)$ . Yet, the probability of a given interval can be very different. Although the relationship between the mean and variance, and the probability of an event is not a direct one, we can still obtain some information about the probabilities based on the mean and variance. In particular, it is possible to bound the probability or to be able to assert that

$$P(|X - E(X)| \geq \varepsilon) \leq \delta < 1.$$

For example, if the probability of a speech signal of mean 0 and variance 1 exceeding a given magnitude  $\varepsilon$  is to be no more than 1%, then we would be satisfied if we could determine a  $\varepsilon$  so that

$$P(|X - E(X)| \geq \varepsilon) \leq \delta < 0.01.$$

We now show that the probability for the event  $|X - E(X)| \geq \varepsilon$  can be bounded if we know the mean and variance. Computation of the probability is not required and therefore the distribution does not need to be known. Estimating the mean and variance is much easier than the entire distribution.

**Theorem 4.1** (Chebyshev inequality). Consider a random variable  $X$ , then we have

$$P(|X - E(X)| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}.$$

Hence, the probability that a random variable deviates from its mean by more than  $\varepsilon$  (in either direction) is less than or equal to  $\frac{\text{Var}(X)}{\varepsilon^2}$ . This agrees with the intuition that the probability of an outcome departing from the mean must become smaller as the width of the distribution decreases or equivalently as the variance decreases.

*Proof.* We have

$$\begin{aligned}
 \text{Var}(X) &= E((X - E(X))^2) = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx \\
 &= \int_{|X-E(X)|>\varepsilon \cup |X-E(X)|\leq\varepsilon} (x - E(X))^2 f_X(x) dx \\
 &\geq \int_{|X-E(X)|>\varepsilon} (x - E(X))^2 f_X(x) dx \\
 &\geq \varepsilon^2 \int_{|X-E(X)|>\varepsilon} f_X(x) dx = \varepsilon^2 P(|X - E(X)| \geq \varepsilon) \\
 \implies P(|X - E(X)| > \varepsilon) &\leq \frac{\text{Var}(X)}{\varepsilon^2}.
 \end{aligned}$$

□

**Theorem 4.2** (Chebyshev inequality and confidence interval). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  with unknown distribution and  $\mu = E(X), \sigma^2 = \text{Var}(X) < \infty$ . Then, for any small positive real number  $\varepsilon$ , we have

$$\begin{aligned}
 P(|\bar{X} - \mu| > \varepsilon) &\leq \frac{\sigma^2}{n\varepsilon^2} \leq \delta \equiv P(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \geq 1 - \delta \\
 \implies \bar{X} &\in \left[ \mu - \sigma \frac{1}{\sqrt{n\delta}}, \mu + \sigma \frac{1}{\sqrt{n\delta}} \right] \text{ with high probability } 1 - \delta.
 \end{aligned}$$

**Example 4.3.** Consider a random variable  $X$  with  $E(X) = 0$  and  $\text{Var}(X) = 1$ , then for an  $\varepsilon$ , we have

$$P(|X| > \varepsilon) \leq \frac{1}{\varepsilon^2}.$$

If  $\varepsilon = 3$ , then

$$P(|X| > 3) \leq \frac{1}{9} \approx 0.11.$$

This is a rather loose bound as follows:

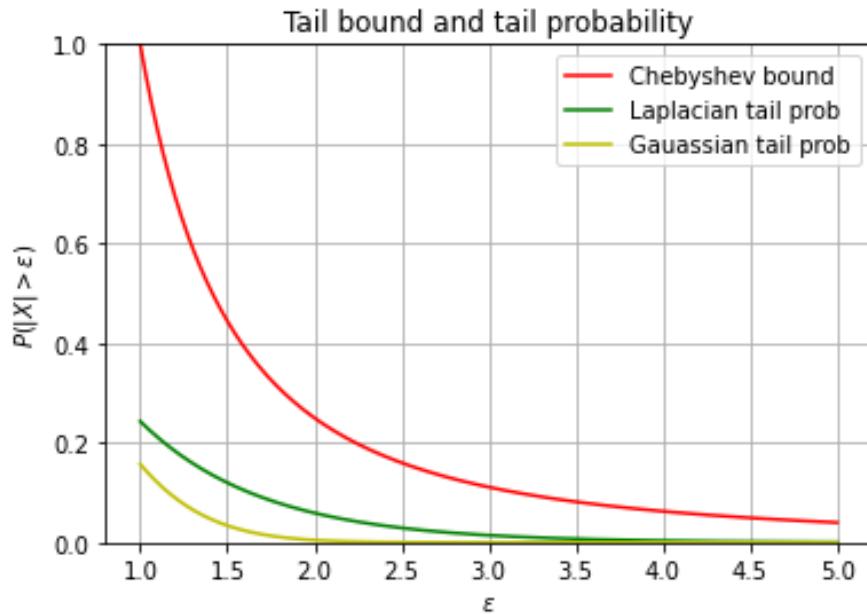
- a. if  $X \sim N(0, 1)$ , then the actual value of this probability is

$$\begin{aligned}
 P(|X| > 3) &= 1 - P(|X| \leq 3) = 1 - (P(3) - P(-3)) \\
 &= 1 - Q(3) + (1 + Q(3)) = 2(1 - Q(3)) \approx 0.0026.
 \end{aligned}$$

b. if  $X \sim L(0, 1)$ , then the actual value of this probability is

$$\begin{aligned} P(|X| > 3) &= \int_{x < -3 \cup x > 3} \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|) dx \\ &= 2 \int_3^\infty \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|) dx \\ &= \exp(-3\sqrt{2}) \approx 0.0144. \end{aligned}$$

Hence, the actual probability is indeed less than or equal to the bound of 0.11, but quite a bit less i.e. the bound provides a gross overestimation of the probability. A graph of the Chebyshev bound as well as the actual probabilities of  $P(|X| > \varepsilon)$  versus  $\varepsilon$  is shown in the below figure.



**Figure 4.1:** Probabilities  $P(|X| > \varepsilon)$  for Gaussian and Laplacian random variables with zero mean and unity variance compared to Ghebyshev inequality.

**Theorem 4.3** (Markov inequality). Consider a non-negative random

variable  $X$ , then we have

$$P(X > \varepsilon) \leq \frac{E(X)}{\varepsilon}.$$

Furthermore, if  $X$  is any random variable and  $g$  is a non-negative function, then

$$P(g(X) > \varepsilon) \leq \frac{E(g(X))}{\varepsilon}.$$

*Proof.* We have

$$\begin{aligned} E(g(X)) &= \int_{-\infty}^{\infty} g(x)f_X(x)dx = \int_{\{x:g(x) \geq \varepsilon \cup g(x) < \varepsilon\}} g(x)f_X(x)dx \\ &\leq \int_{\{x:g(x) \geq \varepsilon\}} g(x)f_X(x)dx \geq \varepsilon \int_{\{x:g(x) \geq \varepsilon\}} f_X(x)dx \\ &= \varepsilon P(g(X) \geq \varepsilon) \implies P(g(X) > \varepsilon) \leq \frac{E(g(X))}{\varepsilon}. \end{aligned}$$

□

**Example 4.4** (Bound for QuickSort run [Tsun, 2020]). Question: Suppose the expected runtime of QuickSort is  $2n \ln(n)$  operations/comparisons to sort an array of size  $n$  (we can show this using linearity of expectation with dependent indicator variables). Use Markov's inequality to bound the probability that QuickSort runs for longer than  $20n \ln(n)$  time.

Answer: Let  $X$  be the runtime of QuickSort, with  $E(X) = 2n \ln(n)$ . Then, since  $X$  is non-negative, we can use Markov's inequality as:

$$E(X > 20n \ln(n)) \leq \frac{E(X)}{20n \ln(n)} = \frac{2n \ln(n)}{20n \ln(n)} = \frac{1}{10}.$$

So we know there's at most 10% probability that QuickSort takes this long to run. Again, we can get this bound despite not knowing anything except its expectation.

**Example 4.5** (coin toss). Question: A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times. Use Markov's inequality to bound the probability it lands on heads at least 16 times.

Answer: We actually do know this distribution; the number of heads is  $X \sim \text{Bin}(n = 20, p = 0.2)$ . Thus,  $E(X) = np = 20 \cdot 0.2 = 4$ . By Markov's inequality: we have

$$P(X \geq 16) \leq \frac{E(X)}{16} = \frac{4}{16} = 0.25, \text{ but}$$

$$P(X \geq 16) = \sum_{k=16}^{20} \binom{n}{k} (0.2)^k (0.8)^{20-k} \approx 1.38 \times 10^{-8}.$$

**Corollary 4.4** (Chebyshev inequality as a corollary of the Markov inequality). Consider a function of  $X$  as  $g(X) = (X - E(X))$ , then we have

$$P(|X - E(X)| \geq \varepsilon) = P((X - E(X))^2 \geq \varepsilon^2) \leq \frac{E(X - E(X))^2}{\varepsilon^2}$$

$$= \frac{\text{Var}(X)}{\varepsilon^2}.$$

**Example 4.6** (coin toss). Question: Consider toss of a weighted coin independently with probability of landing heads  $p = 0.2$ . Upper bound the probability it lands on heads at least 16 times out of 20 flips using Chebyshev's inequality.

Answer: We actually do know this distribution; the number of heads is  $X \sim \text{Bin}(n = 20, p = 0.2)$ . Thus,  $E(X) = np = 20 \cdot 0.2 = 4$ . and variance  $\text{Var}(X) = np(1 - p) = 20 \cdot 0.2 \cdot 0.8 = 3.2$ . By Chebyshev's inequality: we have

$$P(X \geq 16) = P(X - 4 \geq 12) \leq \frac{1}{2} \cdot P(|X - 4| \geq 12) \leq \frac{1}{2} \cdot \frac{\text{Var}(X)}{144}$$

$$= \frac{1}{2} \cdot \frac{3.2}{144} \approx 0.011, \text{ but } P(X \geq 16) = \sum_{k=16}^{20} \binom{n}{k} (0.2)^k (0.8)^{20-k}$$

$$\approx 1.38 \times 10^{-8}.$$

**Remark 4.1.** As we are not able to improve Markov's Inequality and Chebyshev's Inequality in general, it is worth to consider whether we can say something stronger for a more restricted, yet interesting, class of random variables. This idea brings us to consider the case of a random variable that is the sum of a number of independent random variables. This scenario is particularly important and ubiquitous in statistical applications. Examples of such random variables are the number of

heads in a sequence of coin tosses, or the average support obtained by a political candidate in a poll.

Can Markov's and Chebyshev's Inequality be improved for this particular kind of random variable? Before confronting this question, let us check what Chebyshev's Inequality (the stronger of the two) gives us for a sum of independent random variables.

Consider a sequence of random variables  $X_1, X_2, \dots, X_n, \dots$  (i.e.  $\{X_n\}$ ) with  $E(X_i) = \mu_i$  and  $Var(X_i) = \sigma_i^2$ . Then for any real positive number  $\varepsilon$ , we have

$$P\left(\left|\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right| > \varepsilon\right) \leq \frac{Var(\sum_{i=1}^n X_i)}{\varepsilon^2} = \frac{\sum_{i=1}^n \sigma_i^2}{\varepsilon^2}$$

In particular, for identically distributed random variables with expectation  $\mu$  and variance  $\sigma^2$ , we obtain

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

**Theorem 4.5** (Chernoff inequality). Consider a random variable  $X$ , then we have

$$P(|X - E(X)| > \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

Hence, the probability that a random variable deviates from its mean by more than  $\varepsilon$  (in either direction) is less than or equal to  $2 \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$ . This agrees with the intuition that the probability of an outcome departing from the mean must become smaller as the width of the distribution decreases or equivalently as the variance decreases and the sample size increases.

**Remark 4.2.** To improve a point estimation to a high accuracy (i.e. a small confidence interval), we need to improve either assumptions for the tight bound or sample size, because, confidence intervals produced by typical concentration inequalities have size  $O\left(\frac{1}{\sqrt{n}}\right)$ .

**Example 4.7** (Chernoff inequality for a sample mean). Consider a random variable  $X$ , then we have the Chernoff inequality as:

$$P(|X - E(X)| > \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

Now, consider a realization  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim Ber(p)$ . Then, the Chernoff inequality for the sample mean is given by

$$P(|\bar{X} - E(X)| > \varepsilon) \leq 2 \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

**Theorem 4.6** (Chernoff inequality and confidence interval). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a sub-Gaussian random variable  $X$  (i.e.  $M_{X-E(X)}(s) \leq \exp\left(\frac{s^2\sigma^2}{2}\right)$ ) with unknown distribution and  $\mu = E(X), \sigma^2 = Var(X)$ . Then, for any small positive real number  $\varepsilon$ , we have

$$\begin{aligned} P(|\bar{X} - \mu| > \varepsilon) &\leq 2e^{-\frac{n\varepsilon^2}{2\sigma^2}} \leq \delta \equiv P(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - 2e^{-\frac{n\varepsilon^2}{2\sigma^2}} \implies \\ \bar{X} &\in \left[\mu - \sigma\sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}, \mu + \sigma\sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}\right] \text{ with high prob. } 1 - \delta. \end{aligned}$$

**Theorem 4.7** (Chernoff-Hoeffding inequality). Consider a random variable  $X$  that observes values only in  $[a, b]$ , then we have

$$P(|X - E(X)| > \varepsilon) \leq 2 \exp\left(-2\frac{\varepsilon^2}{(b-a)^2}\right).$$

Hence, the probability that a random variable deviates from its mean by more than  $\varepsilon$  (in either direction) is less than or equal to  $\frac{Var(X)}{\varepsilon^2}$ . This agrees with the intuition that the probability of an outcome departing from the mean must become smaller as the width of the distribution decreases or equivalently as the variance decreases.

**Theorem 4.8** (Chernoff-Hoeffding inequality for sample mean). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  having a unknown distribution supported on  $[a, b]$ . Then, for any small positive real number  $\varepsilon > 0$ , we have

$$P(|\bar{X} - \mu| > \varepsilon) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \leq \delta.$$

In particular, when  $X$  is supported on  $[a, b]$  we have

$$P(|\bar{X} - \mu| > \varepsilon) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \leq \delta.$$

**Theorem 4.9** (Chernoff-Hoeffding inequality and confidence interval). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  having an unknown distribution supported on  $[a, b]$ . Then, for any small positive real number  $\varepsilon$ , we have

$$P(|\bar{X} - \mu| > \varepsilon) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \leq \delta \equiv P(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \implies \bar{X} \in \left[ \mu - (b-a)\sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)}, \mu + (b-a)\sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)} \right] \text{ w. h. p. } 1 - \delta.$$

#### 4.2.2 Stochastic convergence and asymptotic large sample theory

**Definition 4.4** (Convergence in probability). Consider a sequence of random variables  $X_1, X_2, \dots, X_n, \dots$  (i.e.  $\{X_n\}$ ). Then,  $\{X_n\}$  converges in probability to a random variable  $X$  for a small positive real number  $\varepsilon$ , if

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0 \equiv \lim_{n \rightarrow \infty} P(|X_n - X| \leq \varepsilon) = 1.$$

This is abbreviated as

$$X_n \rightarrow X \equiv \text{plim}_{n \rightarrow \infty} X_n = X.$$

**Example 4.8** (Sample variance). Consider a random sample of size  $n$  i.e.  $n$  copies of independent and identically distributed random variables  $X_1, X_2, \dots, X_n$  of a random  $X$  with unknown distribution having mean  $\mu = E(X)$  and variance  $\sigma^2 = \text{Var}(X)$ . Then

$$\lim_{n \rightarrow \infty} P(|\bar{X} - E(X)| > \varepsilon) = 0 \equiv \lim_{n \rightarrow \infty} P(|\bar{X} - E(X)| \leq \varepsilon) = 1$$

We have

$$P(|\bar{X} - E(X)| > \varepsilon) \leq \frac{\text{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

**Example 4.9.** Let  $X$  be a random variable such that  $X_n = X + Y_n$  where  $E(Y_n) = \frac{1}{n}$  and  $\text{Var}(Y_n) = \frac{\sigma^2}{n}$ . Then prove that  $X_n$  converges to  $X$  in probability as  $n \rightarrow \infty$ .

Solution: for a random variable  $X$ , it is given that

$$X_n = X + Y_n, \text{ where } E(Y_n) = \frac{1}{n}, \text{ Var}(Y_n) = \frac{\sigma^2}{n}.$$

Now, we have

$$|Y_n| = |Y_n - E(Y_n) + E(Y_n)| \leq |Y_n - E(Y_n)| + \frac{1}{n}. \quad (4.1)$$

Then, using the above inequality (4.9) and Chebyshev inequality, we have

$$\begin{aligned} P(|X_n - X| > \varepsilon) &= P(|Y_n| > \varepsilon) \leq P(|Y_n - E(Y_n)| > \varepsilon) \leq \frac{Var(Y_n)}{\varepsilon} \\ &= \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

**Definition 4.5** (Convergence in mean square). Consider a sequence of random variables  $X_1, X_2, \dots, X_n, \dots$  (i.e.  $\{X_n\}$ ). Then,  $\{X_n\}$  converges in mean square error to a random variable  $X$  for a small positive real number  $\varepsilon$ , if

$$\lim_{n \rightarrow \infty} E((X_n - X)^2) = 0.$$

This is abbreviated as

$$X_n \rightarrow X \equiv \text{mselim}_{n \rightarrow \infty} X_n = X.$$

**Remark 4.3** (Consistent an estimator in mean square error). The 2-mean convergence plays a remarkable role in defining a consistent estimator. Consider a sample statistic  $\hat{\Theta} = T(X_1, X_2, \dots, X_n)$  of a parameter  $\theta$ . We say that the estimator is consistent if its Mean Squared Error (MSE)

$$\begin{aligned} MSE(\hat{\Theta}) &= E((\hat{\Theta} - \theta)^2) = E\left(\left(\hat{\Theta} - E(\hat{\Theta}) + E(\hat{\Theta}) - \theta\right)^2\right) \\ &= E((\hat{\Theta} - E(\hat{\Theta}))^2) + (E(\hat{\Theta}) - \theta)^2 \\ &\quad - 2E(\hat{\Theta} - E(\hat{\Theta}))(E(\hat{\Theta}) - \theta) \\ &= Var(\hat{\Theta}) + Bias^2(\hat{\Theta}) - 0 \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

**Definition 4.6** (Convergence in distribution). Consider a sequence of random variables  $X_1, X_2, \dots, X_n, \dots$  (i.e.  $\{X_n\}$ ). Then,  $\{X_n\}$  converges in distribution to a random variable  $X$  for a small positive real number  $\varepsilon$ , if

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x) \equiv \lim_{n \rightarrow \infty} F_{X_n}(x) = F(x)..$$

This is abbreviated as

$$X_n \rightarrow X \equiv d\lim_{n \rightarrow \infty} X_n = X.$$

**Example 4.10** (Sample mean). Consider a random sample of size  $n$  i.e.  $n$  copies of independent and identically distributed random variables  $X_1, X_2, \dots, X_n$  of a random  $X$  with unknown distribution having mean  $\mu = E(X)$  and variance  $\sigma^2 = \text{Var}(X)$ . Then

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X} - \mu}{\sigma} \leq z\right) = \varphi(z).$$

**Definition 4.7** (Convergence almost surely). Consider a sequence of random variables  $X_1, X_2, \dots, X_n, \dots$  (i.e.  $\{X_n\}$ ). Then,  $\{X_n\}$  converges almost surely to a random variable  $X$  if

$$P\left(\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}\right).$$

A probability of outcomes with point-wise convergence.

**Remark 4.4.** Let  $X_n$  be a sequence of random variables and  $X$  a random variable. Then the following implication diagram is satisfied:

$$\begin{array}{ccccccc} X_n & \xrightarrow{r} & X & \implies & X_n & \xrightarrow{\mathbb{P}} & X \\ & & & & \Downarrow & & \\ & & & & X_n & \xrightarrow{d} & X \end{array}$$

**Figure 4.2:** Relation and implication of various stochastic convergences of a sequence of random variables.

### 4.3 Limit theorems

Often, our goal is to understand how collections of a large number of independent random variables behave. This is what the laws of large numbers reveal. In general, the idea is that the average of a large number of I.I.D. random variables will approach the expectation. Sometimes that is enough, but usually, we also need to understand how fast does the average converge to the expectation.

### 4.3.1 Law of large numbers

**Theorem 4.10** (Strong law of large numbers). Consider a random sample of size  $n$  i.e.  $n$  copies of independent and identically distributed random variables  $X_1, X_2, \dots, X_n$  of a random  $X$  with unknown distribution having mean  $\mu = E(X)$  and variance  $\sigma^2 = Var(X)$ . Then

$$\lim_{n \rightarrow \infty} P(|\bar{X} - E(X)| > \varepsilon) = 0 \equiv \lim_{n \rightarrow \infty} P(|\bar{X} - E(X)| \leq \varepsilon) = 1.$$

*Proof.* Consider the probability of the sample mean random variable deviating from the expected value by more than  $\varepsilon$ , where  $\varepsilon$  is a small positive number. This probability is given by

$$\begin{aligned} P(|\bar{X} - E(\bar{X})| > \varepsilon) &= P(|\bar{X} - E(X)| > \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

It says that for large enough  $n$ , the probability of the error in the approximation of  $\bar{X}$  by  $E(X)$  exceeding  $\varepsilon$  (which can be chosen as small as desired) will be exceedingly small.  $\square$

**Example 4.11** (Bernoulli law of large numbers). Prove that if a fair coin is tossed  $n$  times in succession, then the relative frequency of heads, i.e., the number of heads observed divided by the number of coin tosses, should be close to  $1/2$ . This was why we intuitively accepted the assignment of a probability of  $1/2$  to the event that the outcome of a fair coin toss would be a head.

*Proof.* Consider coin toss experiment as a sequence of  $n$  Bernoulli sub-experiments i.e. a random sample of size  $n$  for the experiment i.e.  $(X_1, X_2, \dots, X_n), X_i \sim Ber(1/2)$ .

, Now, all possible relative frequency is given by the sample mean of the random sample of size  $n$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

In order to see asymptotic behaviour of the sample mean  $\bar{X}$ , we com-

pute the following:

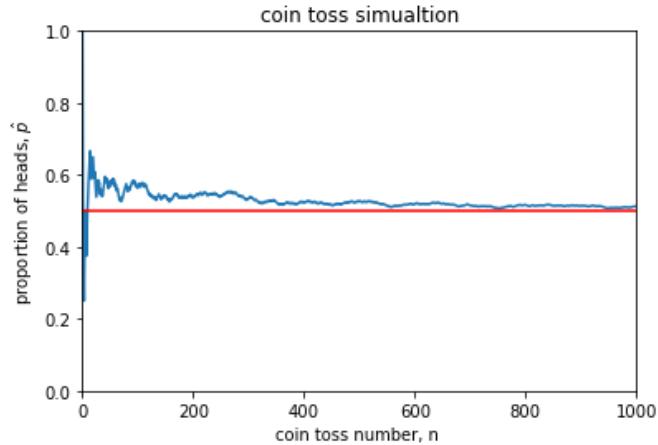
$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n E(X_i)}{n} = 1/2.$$

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n Var(X_i)}{n^2} = \frac{1}{4n}.$$

Now, as per definition of variance, we have

$$Var(\bar{X}) = E((\bar{X} - E(\bar{X}))^2) = \frac{1}{4n} \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\implies \bar{X} \rightarrow E(\bar{X}) = 1/2 \text{ as } n \rightarrow \infty.$$



**Figure 4.3:** A realization of sample mean random variable of  $n$  I.I.D. Bernoulli random variables with  $p = 1/2$  as  $n$  increases.

More generally for a Bernoulli sub-experiment with probability  $p$ , we have that

$$Var(\bar{X}) = E((\bar{X} - E(\bar{X}))^2) = \frac{p(1-p)}{n} \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\implies \bar{X} \rightarrow E(\bar{X}) = p \text{ as } n \rightarrow \infty.$$

□

**Example 4.12** (Sample mean for I.I.D. Gaussian random variables). Consider a sequence of  $n$  Gaussian sub-experiments i.e. a random sample of size  $n$  for the experiment i.e.  $(X_1, X_2, \dots, X_n)$ ,  $X_i \sim N(\mu, \sigma^2)$ . Then  $\bar{X} \rightarrow \mu$  in probability.

### 4.3.2 Limit theorems

At an intuitive level, it says that the appropriately scaled sum of a bunch of independent random variables behaves like a Gaussian (AKA Normal) random variable.

**Theorem 4.11** (Central limit theorem). Consider a random sample of size  $n$  i.e.  $n$  copies of independent and identically distributed random variables  $X_1, X_2, \dots, X_n$  of a random  $X$  with unknown distribution having mean  $\mu = E(X)$  and variance  $\sigma^2 = Var(X)$ . And, define a standardized sample mean is given by

$$\bar{Z}_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X})}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}.$$

Then, for any  $z$ , we have

$$\lim_{n \rightarrow \infty} P(\bar{Z}_n \leq z) = P(Z \leq z) = \varphi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

In summary,  $\bar{Z}_n \rightarrow Z \sim N(0, 1)$  in distribution. Furthermore, the Berry-Esséen inequality gives a more precise quantification of the speed of this convergence

$$\left\| P(\bar{Z}_n \leq z) - \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \right\| \leq \frac{0.77E(|X_1 - E(X_1)|^3)}{(Var(X_1))^{3/2}\sqrt{n}}.$$

*Proof.* Consider a random sample of size  $n$  i.e.  $n$  I.I.D. copies of  $X$  as  $(X_1, X_2, \dots, X_n)$  and the corresponding standardized sample mean is given by

$$\bar{Z}_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} = \frac{n\bar{X}_n - nE(X)}{\sqrt{nVar(\bar{X})}} = \frac{S_n - nE(X)}{\sqrt{nVar(X)}}.$$

Now, we have

$$\begin{aligned}
M_{\bar{Z}_n}(s) &= E_{\bar{Z}_n}(\exp(s\bar{Z}_n)) = E \left( \exp \left( \frac{s}{\sigma\sqrt{n}} (\sum_{i=1}^n X_i - n\mu) \right) \right) \\
&= E \left( \exp \left( \left( \frac{s}{\sqrt{n}} \right) \left( \frac{X_1 - \mu}{\sigma} \right) + \dots + \left( \frac{s}{\sqrt{n}} \right) \left( \frac{X_n - \mu}{\sigma} \right) \right) \right) \\
&= E \left( \exp \left( \frac{s}{\sqrt{n}} \frac{X_1 - \mu}{\sigma} \right) \right) \cdots E \left( \exp \left( \frac{s}{\sqrt{n}} \frac{X_1 - \mu}{\sigma} \right) \right) \\
&= \left( E \left( \exp \left( \frac{s}{\sqrt{n}} \frac{X - \mu}{\sigma} \right) \right) \right)^n = \left( M_{\frac{X-\mu}{\sigma}} \left( \frac{s}{\sqrt{n}} \right) \right)^n \\
&\approx \left( M_{\frac{X-\mu}{\sigma}}(0) + M'_{\frac{X-\mu}{\sigma}}(0) \frac{s}{\sqrt{n}} + M''_{\frac{X-\mu}{\sigma}}(s/\sqrt{n}) \frac{s^2}{2n} \right)^n \\
&\approx \left( 1 + \frac{s^2}{2n} \right)^n \rightarrow \exp \left( \frac{s^2}{2} \right) \text{ as } n \rightarrow \infty \implies \bar{Z}_n \xrightarrow{D} Z \sim \mathcal{N}(0, 1).
\end{aligned}$$

□

**Example 4.13** ([Bertsekas and Tsitsiklis, 2008]). We have 100 bags, with weight distributed uniformly in [5, 50], so that the mean weight is 27.5 lbs and the variance is 168.75lbs. Can we upper bound the probability that the total weight is larger than 3000lbs?

For comparison, let's see what happens if we use Chebyshev's inequality

$$\begin{aligned}
P(S_{100} > 3000) &= \frac{1}{2} \times P \left( \left| \frac{1}{100} \sum_{i=1}^{100} X_i - \frac{2750}{100} \right| > 2.5 \right) \\
&\leq \frac{1}{2} \times \frac{\text{Var}(X)}{100 \cdot (2.5)^2} \approx 0.13 = 13\%.
\end{aligned}$$

That is a strict inequality, but intuitively it is overestimating the probability by a factor of two since it is also including the case of the average being significantly smaller than the mean as well.

If we use the CLT instead,

$$\begin{aligned}
 P\left(\sum_{i=1}^{100} X_i > 3000\right) &= 1 - P\left(\sum_{i=1}^{100} X_i \leq 3000\right) \\
 &= 1 - P\left(\frac{\sum_{i=1}^{100} X_i - 3000}{\sqrt{100 \cdot 168.75}} \leq \frac{275}{\sqrt{100 \cdot 168.75}}\right) \\
 &\approx 1 - P\left(Z \leq \frac{275}{\sqrt{100 \cdot 168.75}}\right) = 0.027.
 \end{aligned}$$

where  $Z$  is a random variable distributed as  $N(0, 1)$ . So the CLT predicts that the probability of having a total weight of over 3000lbs is about 3%, which is much lower than the 27% bound (or 13% if we divide by two) that we got from Chebyshev's inequality! It turns out that the Normal approximation here is very good even though the CLT is not a bound and 3% is not an appropriately low number according to the Berry-Esséen inequality. This is because the CLT is very good for the sums of bounded random variables like the uniform.

**Remark 4.5** (An inequality for computing tail bounds for Gaussian distribution). In order to use the CLT to get easily calculated bounds, the following approximations will often prove useful: for any  $z > 0$ ,

$$\left(1 - \frac{1}{z^2}\right) \frac{\exp\left(-\frac{z^2}{2}\right)}{z\sqrt{2\pi}} \leq \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \frac{\exp\left(-\frac{z^2}{2}\right)}{z\sqrt{2\pi}}.$$

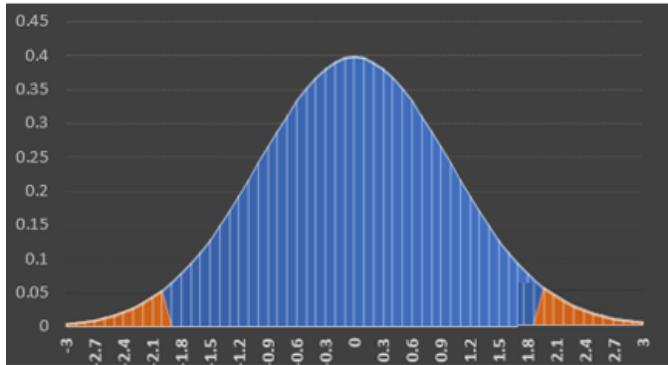
**Example 4.14.** Backdrop: Assume a realization/people  $\mathbf{x} = (x_1, \dots, x_n)$  (of an I.I.D random sample of size  $n$  from  $X \sim Ber(p)$ ) vote party A with probability  $p$ . We poll  $n$  people. Let  $M_n$  be the fraction/average of the polled persons for Party A and defined as

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Find  $n$  such that estimator  $M_n$  to be within  $\pm 1\%$  of the true value  $p$  with probability at least 95%, statistically

$$P(|M_n - p| > 0.01) \leq 0.05.$$

And geometrically, can be visualized as in the figure (4.4):



**Figure 4.4:** The distribution  $M_n - p$  is symmetric, and we are looking for tail bounds.

Answer: Since the distribution of  $M_n$  is symmetric around its mean, we can use this symmetry to write  $P(|M_n - p| > 0.01) \approx 2P(M_n - p > 0.01)$ . Because the  $X_i \sim Ber(p)$ , we can bound their variance by  $1/4$ , so  $Var(M_n) \leq 1/4n$ . Dividing both sides by the square root of the variance to put the probability into the CLT form, we get

$$\begin{aligned} P(|M_n - p| > 0.01) &\approx 2P(M_n - p > 0.01) \approx 2P\left(\frac{M_n - p}{\sqrt{1/4n}} > \frac{0.01}{\sqrt{1/4n}}\right) \rightarrow \\ 2P\left(Z > \frac{0.01}{\sqrt{1/4n}}\right) &\Rightarrow P\left(Z > \frac{0.01}{\sqrt{1/4n}}\right) = 1 - \varphi\left(\frac{0.01}{\sqrt{1/4n}}\right) \leq .025 \\ \Rightarrow \varphi\left(\frac{0.01}{\sqrt{1/4n}}\right) &\geq 0.9750 \Rightarrow \frac{0.01}{\sqrt{1/4n}} \geq 1.96 \Rightarrow n \geq 9604. \end{aligned}$$

In this calculation we used the approximation given by the CLT, but how good is it? If we work out the exact computation, using the fact that  $M_n$  is a scaled binomial and the worse-case parameter  $p = 1/2$ , then we get that  $n = 9604$  will give us a 95.1% probability of being within 0.01 of the true value  $p$ : in this case, the CLT's approximation is very good. So if the CLT is so much better, why use Chebyshev's inequality at all? The reason is that Chebyshev is a bound that works for all  $\varepsilon$  and probabilities, is always a valid upper bound but is very conservative. The CLT is often much sharper, however it only works for  $\varepsilon$

of appropriate scale and does not give a rock-solid bound as it is an approximation and not a bound. Is it possible to strengthen Chebyshev's inequality to get a bound nearly as sharp as the CLT's, under appropriate assumptions? Yes, the Chernoff inequality provides the optimal approximation.

**Remark 4.6** (Chernoff inequality and CLT). To get some insight into the CLT by exploring how the value of the KL divergence  $D(a\|p)$  of a scales when  $a$  is in the vicinity of  $p$ .

$$D(a\|p) = E_a \left( \ln \left( \frac{a}{p} \right) \right) =$$

$$\begin{cases} \sum_{x \in \Omega_X} \ln \left( \frac{a(x)}{p(x)} \right) a(x) = \sum_{x \in \Omega_X} a(x) \ln(a(x)) - \sum_{x \in \Omega_X} a(x) \ln(p(x)) \\ \int_{x \in \Omega_X} \ln \left( \frac{a(x)}{p(x)} \right) a(x) dx = \int_{x \in \Omega_X} a(x) \ln(a(x)) dx - \int_{x \in \Omega_X} a(x) \ln(p(x)) dx. \end{cases}$$

KL divergence  $D(a\|p)$  of  $a$  in the vicinity of  $p$  is the difference of entropy of  $a$  and expected value of  $\ln(p)$  with respect to  $a$ :

$$D(a\|p) = a \ln(a) + (1-a) \ln(1-a) - a \ln(p) - (1-a) \ln(1-p).$$

# 5

---

## Statistical methods

---

In the previous chapter we discuss large sample theory that plays very important role in the statistical inference of statistical methods to be discussed here in details.

### 5.1 Point estimation

There are various approaches of point estimations such as

- (i) maximum likelihood estimation i.e. MLE
- (ii) method of moments
- (iii) Max a posterior estimation i.e. MAP.

#### 5.1.1 Sample statistic

Consider a random variable (or a measurement)  $X$  with unknown distribution, then it is very difficult to compute statistic or parameter of the observed values under the unknown distribution. Now, we need to define sample statistic instead with respect to a random sample of size  $n$  (i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  independent and identically distributed

(i.i.d.) copies of  $X$ ). Then, recalling concepts of sample statistic from subsection (??) of chapter (4):

- a. Sample mean:

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} \text{ with a realization } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

- b. Sample variance:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

with a realization  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$

- c. Sample standard deviation:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}, \text{ with a realization}$$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

### 5.1.2 Maximum likelihood estimation

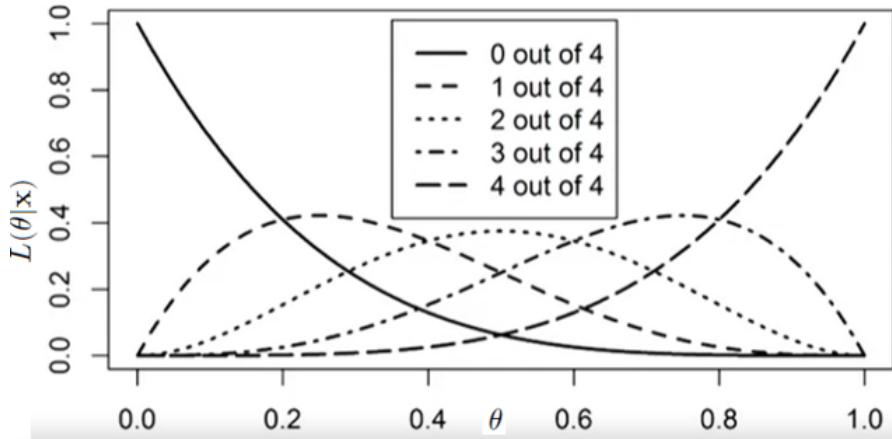
#### Likelihood function

**Definition 5.1.** Consider a realization/data  $\mathbf{x} = (x_1, \dots, x_n)$  an i.i.d. random sample of size  $n$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , where  $\theta$  is a parameter (or vector of parameters). Now, we define a likelihood of  $\mathbf{x}$  given  $\theta$  to be the probability of observing the data  $\mathbf{x}$  as

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \begin{cases} \prod_{i=1}^n p_X(x_i|\theta) & \text{if } X \text{ is discrete} \\ \prod_{i=1}^n f_X(x_i|\theta) & \text{if } X \text{ is continuous.} \end{cases}$$

The likelihood function  $L(\theta|\mathbf{x})$  has the following silent features:

- a. The likelihood function is not a probability density function.



**Figure 5.1:** A likelihood function  $L(\theta|\mathbf{x})$  wrt a probabilistic model  $p_{\mathbf{X}}(x|\theta)$  or  $f_{\mathbf{X}}(x|\theta)$  with a unknown parameter  $\theta$ .

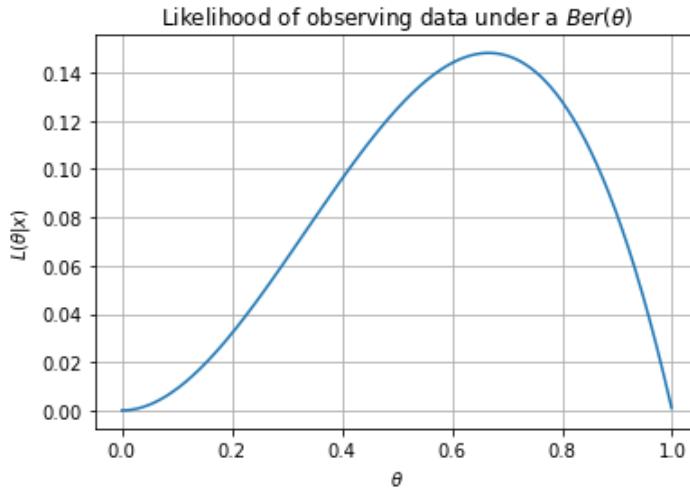
- b. It is an important component of both frequentist and Bayesian analyses.
- c. It measures the support provided by the data for each possible value of the parameter.

If we compare the likelihood function at two parameter points and find that  $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$  then the realized sample data  $\mathbf{x}$ , we actually observed, is more likely to have occurred if  $\theta = \theta_1$  than if  $\theta = \theta_2$ . This can be interpreted as  $\theta_1$  is a more plausible value for  $\theta$  than  $\theta_2$ .

**Example 5.1** (A Bernoulli model). Question: Consider a realization  $\mathbf{x} = (x_1 = 1, x_2 = 0, x_3 = 1)$  of an i.i.d. sample of size 3 from  $X \sim Ber(\theta)$ , where  $\theta$  is the unknown probability of a success. Compute the likelihood function  $L(\theta|\mathbf{x})$ .

Answer: it is given a realization  $\mathbf{x} = (x_1 = 1, x_2 = 0, x_3 = 1)$  of an i.i.d. samples from  $X \sim Ber(\theta)$ , then we have

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^3 p_X(x_i|\theta) = \prod_{i=1}^3 \theta^{x_i} (1-\theta)^{1-x_i} = \theta^2 (1-\theta).$$



**Figure 5.2:** A likelihood function  $L(\theta|x) = \theta^2(1 - \theta)$  wrt  $X \sim Ber(\theta)$  with a unknown parameter  $\theta$ .

**Example 5.2.** Consider a realization  $\mathbf{x} = (x_1 = 3, x_2 = 0, x_3 = 2, x_4 = 7)$  of an I.I.D. sample of size 4 from  $X \sim Pois(\theta)$ , where  $\theta$  is the historical average number of events in a unit of time).

Answer: it is given a realization  $\mathbf{x} = (x_1 = 1, x_2 = 0, x_3 = 1)$  of an I.I.D. samples from  $X \sim Ber(\theta)$ , then we have

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^4 p_X(x_i|\theta) = \prod_{i=1}^4 e^\theta \frac{\theta^{x_i}}{(x_i)!} = e^{-4\theta} \frac{\theta^{12}}{2!3!7!}.$$

### Maximum likelihood estimation

**Definition 5.2** (Maximum likelihood estimate). Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , where  $\theta$  is a parameter (or vector of parameters). Then, the maximum likelihood estimate  $\hat{\theta}_{MLE}$  of  $\theta$  to be the parameter which maximizes the likelihood (or equivalently, the log-likelihood) of the data as:

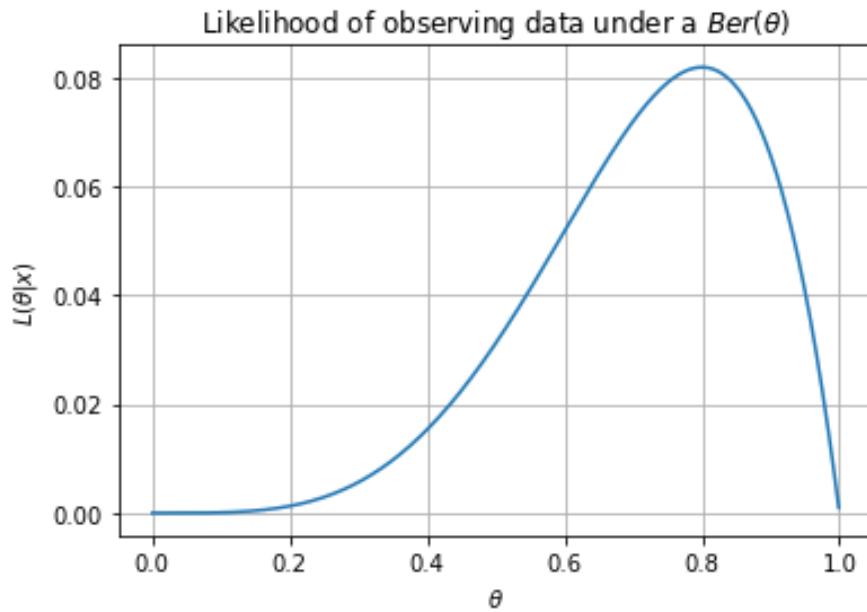
$$\hat{\theta} = \arg \max_{\theta} L(\theta|\mathbf{x}) = \arg \max_{\theta} \ln(L(\theta|\mathbf{x})).$$

**Example 5.3** (Bernoulli model). Question: consider a realization  $\mathbf{x} = (x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0)$  of an I.I.D. sample of size 4 from  $X \sim Ber(\theta)$  distribution with unknown parameter  $\theta$ . Find the maximum likelihood estimate  $\hat{\theta}$  of  $\theta$ .

We have the likelihood function with respect to the given realization of the sample as

$$L(\theta|\mathbf{x}) = \prod_{i=1}^5 p_X(x_i|\theta) = \prod_{i=1}^5 \theta^{x_i} (1-\theta)^{1-x_i} = \theta^4(1-\theta).$$

The plot of the likelihood with  $\theta$  on the x-axis and  $L(HHHHT|\theta)$  on the y-axis in (5.3) as:



**Figure 5.3:** A likelihood function with the Bernoulli model for unknown parameter  $\theta$  (the probability of heads).

and we can actually see the  $\theta$  which maximizes the likelihood is  $\hat{\theta} = 4/5$ . But sometimes it is not possible to plot the likelihood, so we solve it analytically now.

To find the  $\theta$  which maximizes this likelihood, so we take the derivative with respect to  $\theta$  and set it to 0:

$$\frac{d}{d\theta}(L(\theta|\mathbf{x})) = \frac{d}{d\theta}(\theta^4(1-\theta)) = \theta^3(4-5\theta) = 0 \implies \theta_{opt} = 4/5.$$

So, in MLE, we're trying to find the  $\theta$  that maximizes the likelihood, and we don't care what the maximum value of the likelihood is. We didn't even compute it. We just care that the  $\arg \max_{\theta} L(\theta|x)$  is  $4/5$ .

**Remark 5.1** (Likelihood vs log-likelihood). Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , where  $\theta$  is a parameter (or vector of parameters). The log-likelihood of  $\mathbf{x}$  given  $\theta$  are defined as:

$$\ln(L(\theta|x)) = \begin{cases} \sum_{i=1}^n \ln(p_X(x_i|\theta)) & \text{X is discrete} \\ \sum_{i=1}^n \ln(f_X(x_i|\theta)) & \text{X is continuous.} \end{cases}.$$

Since,  $\ln(\cdot)$  is a monotone increasing function, so it preserves order of the argument, so whatever was the maximizer ( $\arg \max$ ) in the original function, will also be maximizer in the log function. For example, we can see in the following picture

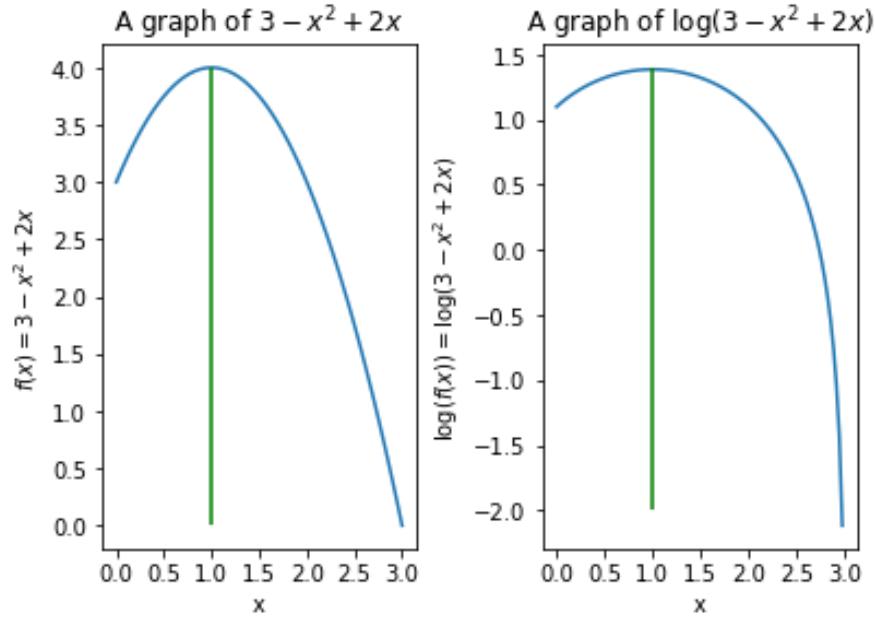
**Example 5.4.** Consider a realization  $\mathbf{x} = (x_1 = 1, x_2 = 0, x_3 = 1)$  of an I.I.D. sample of size  $n = 3$  from  $X \sim Ber(\theta)$ . Compute the MLE estimate of  $\theta$ .

Answer: Since  $X \sim Ber(\theta)$ , so we have  $p_X(x|\lambda) = \lambda^x(1-\lambda)^{1-x}$ . Now, likelihood function is defined as

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}.$$

Taking log both sides, we get

$$\begin{aligned} \ln(L(\theta|\mathbf{x})) &= \sum_{i=1}^n (x_i \ln(\theta) + (1-x_i) \ln(1-\theta)) \\ &= \ln(\theta) \sum_{i=1}^n x_i + \ln(1-\theta) \sum_{i=1}^n (1-x_i). \end{aligned}$$



**Figure 5.4:** The values are different in y-axis, but if you look at the x-axis, it happens that both functions are maximized at 1 i.e. the arg max's are the same.

Take the derivative(s) with respect to  $\theta$  and set to 0, we have

$$\begin{aligned}
 \frac{d}{d\theta} (\ln(L(\theta|x))) &= \frac{1}{\theta} \sum_{i=1}^n x_i - \frac{1}{1-\theta} \sum_{i=1}^n (1-x_i) \\
 &= \frac{1}{\theta(1-\theta)} \left( (1-\theta) \sum_{i=1}^n x_i - \theta \sum_{i=1}^n (1-x_i) \right) \\
 &= \frac{1}{\theta(1-\theta)} \left( \sum_{i=1}^n x_i - n\theta \right) \\
 \frac{d^2}{d\theta^2} (\ln(L(\theta|x))) &= \frac{-1}{\theta^2} \sum_{i=1}^n x_i - \frac{1}{(1-\theta)^2} \sum_{i=1}^n (1-x_i)
 \end{aligned}$$

Equating the derivative of  $\ln(L(\theta|x))$  to 0, we get

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \text{ with } \frac{d^2}{d\theta^2} (\ln(L(\theta|x))) \leq 0. \text{ So, } \hat{\Theta} = \bar{X}.$$

**Example 5.5** (Poisson model). Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from  $X \sim Pois(\theta)$ . Compute the MLE estimate of  $\theta$ .

Answer: Since  $X \sim Ber(\theta)$ , so we have  $p_X(x|\theta) = e^{-\theta} \frac{\theta^x}{x!}$ . Now, likelihood function is defined as:

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{(x_i)!}.$$

Now, taking log sides, we get

$$\begin{aligned} \ln(L(\theta|\mathbf{x})) &= \ln(\prod_{i=1}^n p_X(x_i|\theta)) = \sum_{i=1}^n \ln(p_X(x_i|\theta)) \\ &= \sum_{i=1}^n -\theta + x_i \ln(\theta) - \ln((x_i)!). \end{aligned}$$

Take the derivative(s) with respect to  $\theta$  and set to 0, we have

$$\begin{aligned} \frac{d}{d\theta} \ln(L(\theta|\mathbf{x})) &= \sum_{i=1}^n \left( -1 + \frac{x_i}{\theta} \right) = 0 \implies -n + \frac{1}{\theta} \sum_{i=1}^n x_i = 0 \implies \\ \hat{\theta} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \text{ with } \frac{d^2}{d\theta^2} \ln(L(\theta|\mathbf{x})) = \sum_{i=1}^n \left( -\frac{x_i}{\theta^2} \right) \leq 0. \text{ So, } \hat{\Theta} = \bar{X}. \end{aligned}$$

**Example 5.6.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from  $X \sim Exp(\theta)$ . Compute the MLE estimate of  $\theta$ .

Answer: Since  $X \sim Ber(\theta)$ , so we have  $p_X(x|\theta) = \theta e^{-x\theta}$ . Now, likelihood function is defined as:

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n \theta e^{-x_i\theta}.$$

Now, taking log sides, we get

$$\begin{aligned} \ln(L(\theta|\mathbf{x})) &= \ln(\prod_{i=1}^n p_X(x_i|\theta)) = \sum_{i=1}^n \ln(p_X(x_i|\theta)) \\ &= \sum_{i=1}^n -x_i\theta + \ln(\theta). \end{aligned}$$

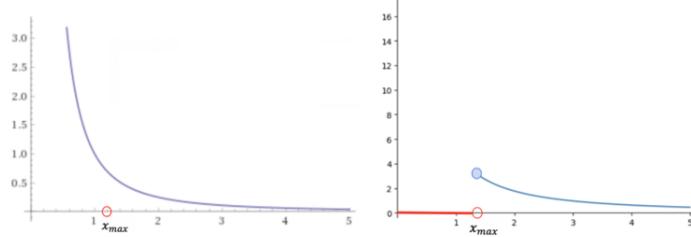
Take the derivative(s) with respect to  $\theta$  and set to 0, we have

$$\begin{aligned} \frac{d}{d\theta} \ln(L(\theta|x)) &= \sum_{i=1}^n \left( -x_i + \frac{1}{\theta} \right) = 0 \implies \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \implies \\ \hat{\theta} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \text{ with } \frac{d^2}{d\theta^2} \ln(L(\theta|x)) = \sum_{i=1}^n \left( -\frac{1}{\theta^2} \right) \leq 0. \text{ So, } \hat{\Theta} = \bar{X}. \end{aligned}$$

**Example 5.7.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from  $X \sim \text{Unif}(0, \theta)$ . Compute the MLE estimate of  $\theta$ .

Answer: Since  $X \sim \text{Ber}(\theta)$ , so we have  $p_X(x|\theta) = \frac{1}{\theta} I_{0 \leq x \leq \theta}$ . Now, likelihood function is defined as:

$$\begin{aligned} L(x|\theta) &= \prod_{i=1}^n p_X(x_i|\theta) = p_X(x_1|\theta) \cdot p_X(x_2|\theta) \cdots p_X(x_n|\theta) \\ &= \frac{1}{\theta} I_{\{0 \leq x_1 \leq \theta\}} \cdot \frac{1}{\theta} I_{\{0 \leq x_2 \leq \theta\}} \cdots \frac{1}{\theta} I_{\{0 \leq x_n \leq \theta\}} \\ &= \frac{1}{\theta^n} I_{\{0 \leq x_1, x_2, \dots, x_n \leq \theta\}}. \end{aligned}$$



**Figure 5.5:** The first (left) plot is graph of  $\frac{1}{\theta^n}$ , and the second (right) plot of the actual likelihood function  $\frac{1}{\theta^n} I_{\{0 \leq x_1, \dots, x_n \leq \theta\}} \equiv \frac{1}{\theta^n} I_{\{x_{max} \leq \theta\}}$ .

Take the derivative(s) with respect to  $\theta$  and set to 0, we have

$$\begin{aligned} \frac{d}{d\theta} L(\theta|x) &= \frac{-n}{\theta^{n+1}} I_{\{0 \leq x_1, x_2, \dots, x_n \leq \theta\}} = 0 \implies \hat{\theta} = \max_{\{i=1,2,\dots,n\}} x_i \\ &= x_{max} \text{ with } \frac{d^2}{d\theta^2} L(\theta|x) = \frac{n(n+1)}{\theta^{n+2}} I_{\{0 \leq x_1, x_2, \dots, x_n \leq \theta\}} \geq 0. \\ \text{So, } \hat{\Theta} &= \max_{\{i=1,2,\dots,n\}} X_i = X_{max}. \end{aligned}$$

### 5.1.3 Method of moments

In the last sub-section, we have seen maximum likelihood estimation (MLE) had a nice intuition but mathematically is a bit tedious to solve. So, we'll learn an another technique for estimating parameters called the Method of Moments (MoM) with several examples which hopefully makes things clearer.

**Definition 5.3** (kth moment). Let  $X$  be a random variable and  $c \in \mathbb{R}$  be a scalar. Then: (a.) the kth moment of  $X$  and (b.) the kth moment of  $X$  (about  $c$ ) are defined as:

$$(a.) E(X^k) \text{ and } (b.) E((X - c)^k) \text{ respectively.}$$

Usually, we are interested in the first moment of  $X$ :  $\mu = E(X)$ , and the second moment of  $X$  about  $\mu$ :  $Var(X) = E((X - \mu)^2)$ .

**Definition 5.4** (kth sample moment). Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$  with unknown parameter  $\theta$  and  $c \in \mathbb{R}$  be a scalar. Then: (a.) the kth sample moment of  $X$  and (b.) the kth sample moment of  $X$  (about  $c$ ) are defined as:

$$(a.) \frac{1}{n} \sum_{i=1}^n x_i^k \text{ and } (b.) \frac{1}{n} \sum_{i=1}^n (X - c)^k \text{ respectively.}$$

Usually, we are interested in the first sample moment of  $X$ :  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , and the second sample moment of  $X$  about  $\mu$ :  $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ .

**Example 5.8.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample size  $n$  from  $X \sim Unif(0, \theta)$ , where  $\theta$  is a unknown parameter. Compute MoM estimator of  $\theta$ .

Answer: since  $X \sim Ber(\theta)$ , so we have  $p_X(x|\theta) = \frac{1}{\theta} I_{0 \leq x \leq \theta}$ , so  $E(X) = \frac{\theta}{2}$ . Then, from first moment approximation, we have

$$E(X) = \frac{\theta}{2} = \frac{1}{n} \sum_{i=1}^n \implies \hat{\theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n x_i.$$

It can be observed that

$$\hat{\theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n x_i \neq \hat{\theta}_{MLE} = x_{max}.$$

And the MoM estimator of  $\theta$  is given by

$$\hat{\Theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n X_i = 2\bar{X}.$$

**Example 5.9.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from  $X \sim Exp(\theta)$ . Compute the MoM estimator of  $\theta$ .

Answer: since  $X \sim Exp(\theta)$ , so we have  $f_X(x|\theta) = \theta e^{-x\theta}$ , so  $E(X) = \frac{1}{\theta}$ . Then, from first moment approximation, we have We set the first true moment to the first sample moment as follows (recall that  $X \sim Exp(\lambda)$ ,  $E(X) = \frac{1}{\lambda}$ ):

$$E(X) = \frac{1}{\theta} = \frac{1}{n} \sum_{i=1}^n x_i \implies \hat{\theta}_{MoM} = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i} = \hat{\theta}_{MLE} = \frac{1}{\bar{x}}.$$

So, the MoM estimator of the  $\theta$  is given by

$$\hat{\Theta}_{MoM} = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i} = \frac{1}{\bar{X}}.$$

**Example 5.10.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample size  $n$  from  $X \sim Pois(\theta)$ . Compute MoM estimator of  $\theta$ .

Answer: since  $X \sim Ber(\theta)$ , so we have  $p_X(x|\theta) = \frac{1}{\theta} I_{0 \leq x \leq \theta}$ , so  $E(X) = \theta$ . Then, from first moment approximation, we have

$$E(X) = \theta = \frac{1}{n} \sum_{i=1}^n \implies \hat{\theta}_{MoM} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Notice that in this case, the MoM estimate disagrees with the MLE as follows:

$$\frac{1}{n} \sum_{i=1}^n x_i = \hat{\theta}_{MoM} = \hat{\theta}_{MLE} = \bar{x}.$$

So, the MoM estimator of  $\theta$  is given by

$$\hat{\Theta}_{MoM} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

### 5.1.4 Max a posterior estimation

In order to begin with Bayesian estimation, it requires us to learn at least few special distributions (e.g. Beta distribution) especially meant for prior distribution assumption.

#### The Beta Random Variable

In order to proceed in Bayesian framework, it is very much essential to define few useful distributions to be utilized for prior distribution of the unknown parameter  $\theta$ .

**Definition 5.5** (Beta distribution [Tsun, 2020]). A random variable X is having beta distribution i.e.  $X \sim Beta(\alpha, \beta)$ , if and only if X has the following density function (and range  $\Omega_X = [0, 1]$ ):

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we've seen  $\alpha - 1$  successes and  $\beta - 1$  failures. Hence the mode (most likely value of the probability/point with highest density)  $\arg \max_{x \in [0,1]} f_X(x)$ , is

$$\text{mode}(X) = \frac{\alpha - 1}{\alpha - 1 + \beta - 1}.$$

Also note the following:

- a. The first term in the pdf,  $\frac{1}{B(\alpha, \beta)}$  is just a normalizing constant (ensures the pdf to integrates to 1). It is called the Beta function, and so our random variable is named as a Beta random variable.
- b. There is an annoying off-by-1 issue: ( $\alpha - 1$  heads and  $\beta - 1$  tails), so when choosing these parameters, be careful.
- c. x is the probability of success, and  $(1 - x)$  is the probability of failure.

**Example 5.11** ([Tsun, 2020]). If you flip a coin with unknown probability of heads X, identify the parameters of the most appropriate Beta distribution to model your belief:

- a. You didn't observe anything (i.e.  $\alpha = \#heads + 1 = 0 + 1 = 1, \beta = \#tails + 1 = 0 + 1 = 1$ )?

It's  $B(0 + 1, 0 + 1) \equiv B(1, 1) \equiv Unif(0, 1) \rightarrow$  NO mode (because it follows the Uniform distribution, every point has same density).

- b. You observed 8 heads and 2 tails?

It's  $B(8 + 1, 2 + 1) \equiv B(9, 3) \rightarrow \text{mode}(X) = \frac{9-1}{9-1+3-1} = \frac{8}{10} = 0.8.$

- c. You observed 80 heads and 20 tails?

It's  $B(80 + 1, 20 + 1) \equiv B(81, 21) \rightarrow \text{mode}(X) = \frac{81-1}{81-1+21-1} = \frac{80}{100} = 0.8.$

- d. You observed 2 heads and 3 tails?

It's  $B(2 + 1, 3 + 1) \equiv B(3, 4) \rightarrow \text{mode}(X) = \frac{3-1}{3-1+4-1} = \frac{2}{5} = 0.4.$

The Dirichlet random vector generalizes the Beta random variable to having a belief distribution over  $(p_1, p_2, \dots, p_r)$  (like in the multinomial distribution so  $\sum_{i=1}^r p_i = 1$ ), and has r parameters  $(\alpha_1, \alpha_2, \dots, \alpha_r)$ . It has the similar interpretation of pretending you've seen  $\alpha_i - 1$  outcomes of type i.

**Definition 5.6** (Dirichlet distribution). A random variable X is having Dirichlet distribution i.e.  $X \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_r)$ , if and only if X has the following density function (and range  $\Omega_X = [0, 1]$ ):

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha_1, \alpha_2, \dots, \alpha_r)} \prod_{i=1}^r x_i^{\alpha_i-1} & x_i \in [0, 1], \sum_{i=1}^r x_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

This is a generalization of the Beta random variable from 2 outcomes to r. X is typically the belief distribution about some unknown probability of different outcomes, where we pretend we've seen  $\alpha_1 - 1$  outcomes of type 1,  $\alpha_2 - 1$  outcomes of type 2 ...,  $\alpha_r - 1$  outcomes of type r. Hence the mode (most likely value of the probability/point with highest density)  $\arg \max_{x_i \in [0, 1], \sum_{i=1}^r x_i = 1} f_X(x)$ , is

$$\text{mode}(X) = \left( \frac{\alpha_1 - 1}{\sum_{i=1}^r (\alpha_i - 1)}, \frac{\alpha_2 - 1}{\sum_{i=1}^r (\alpha_i - 1)}, \dots, \frac{\alpha_r - 1}{\sum_{i=1}^r (\alpha_i - 1)} \right).$$

Also note the following:

- a. Similar to the Beta RV, the first term in the pdf,  $\frac{1}{B(\alpha)}$  is just a normalizing constant (ensures the pdf to integrates to 1), where  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_r)$ .
- b. Notice that this is the probability distribution over the random vector  $x_i$ 's, which is the vector of probabilities, so they must sum to 1 i.e.  $\sum_{i=1}^r x_i = 1$ .

### MAP

**Definition 5.7 (MAP).** Consider a realization/data  $x = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from probability mass function  $p_X(x : \Theta = \theta)$  (if X discrete), or from density  $f_X(x; \Theta = \theta)$  (if X continuous), where  $\Theta$  is the random variable representing the parameter (or vector of parameters). We define the Maximum A Posteriori (MAP) estimate  $\hat{\theta}_{MAP}$  of  $\theta$  as

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \pi_{\Theta}(\theta|x) = \arg \max_{\theta} \pi_{\Theta}(\theta)L_{\Theta}(\theta|x).$$

That is, it's exactly the same as maximum likelihood, except instead of just maximizing the likelihood, we are maximizing the likelihood multiplied by the prior.

**Example 5.12** (uniform prior distribution and  $\hat{\theta}_{MLE} = \hat{\theta}_{MAP}$ ). Question: consider a realization  $\mathbf{x} = (0, 0, 1, 1, 0)$  of an I.I.D. random sample of size 5 from  $X \sim Ber(\Theta)$ , where  $\Theta \in \{0.2, 0.5, 0.7\}$ . Compute the (a.)  $\hat{\theta}_{MLE}$  and  $\hat{\theta}_{MAP}$  with respect to prior distribution  $\pi_{\Theta}(\theta) = 1/3$  for  $\theta \in \{0.2, 0.5, 0.7\}$ .

Solution: it is given a realization  $\mathbf{x} = (0, 0, 1, 1, 0)$  of an i.i.d. sample of size 5 from  $X \sim Ber(\theta)$ , then we have

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^5 p_X(x_i|\theta) = \prod_{i=1}^5 \theta^{x_i} (1-\theta)^{1-x_i} = \theta^2(1-\theta)^3.$$

Then,

$$\begin{aligned} L(0.2|\mathbf{x}) &= P(\mathbf{x}|0.2) = (0.2)^2 \cdot (0.8)^3 = 0.02048, \\ L(0.5|\mathbf{x}) &= P(\mathbf{x}|0.5) = (0.5)^2 \cdot (0.5)^3 = 0.03125, \\ L(0.7|\mathbf{x}) &= P(\mathbf{x}|0.7) = (0.7)^2 \cdot (0.3)^3 = 0.01323. \end{aligned}$$

So,  $\hat{\theta}_{MLE} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(\theta|x) = 0.5$ . Furthermore,

$$\begin{aligned}\pi_\Theta(0.2)L(0.2|x) &= \frac{1}{3} \cdot (0.2)^2 \cdot (0.8)^3 \approx 0.00683, \\ \pi_\Theta(0.5)L(0.5|x) &= \frac{1}{3} \cdot (0.5)^2 \cdot (0.5)^3 \approx 0.01042, \\ \pi_\Theta(0.7)L(0.7|x) &= \frac{1}{3} \cdot (0.7) \cdot (0.3)^3 = 0.00441.\end{aligned}$$

So,  $\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} \pi_\Theta(\theta)L(\theta|x) = 0.5$ . i.e. with a uniform prior distribution of  $\Theta$ , we have

$$\hat{\theta}_{MLE} = 0.5 = \hat{\theta}_{MAP}.$$

**Example 5.13** (Non-uniform prior distribution and  $\hat{\theta}_{MLE} \neq \hat{\theta}_{MAP}$ ).

Question: consider a realization  $\mathbf{x} = (x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0)$  of an I.I.D. random sample of size 4 from  $X \sim Ber(\Theta)$ , where  $\Theta \in \{0.2, 0.5, 0.7\}$  with a prior distribution  $p_\Theta(0.2) = 0.10, p_\Theta(0.5) = 0.01$  and  $p_\Theta(0.7) = 0.89$ . Compute the (a.)  $\hat{\theta}_{MLE}$  and  $\hat{\theta}_{MAP}$ .

Solution: it is given a realization  $\mathbf{x} = (0, 0, 1, 0)$  of an i.i.d. sample of size 4 from  $X \sim Ber(\theta)$ , then we have

$$L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta) = \prod_{i=1}^4 p_X(x_i|\theta) = \prod_{i=1}^4 \theta^{x_i} (1-\theta)^{1-x_i} = \theta(1-\theta)^3.$$

Then,

$$\begin{aligned}L(0.2|x) &= P(x|0.2) = (0.2) \cdot (0.8)^3 = 0.01024, \\ L(0.5|x) &= P(x|0.5) = (0.5) \cdot (0.5)^3 = 0.06250, \\ L(0.7|x) &= P(x|0.7) = (0.7) \cdot (0.3)^3 = 0.01890.\end{aligned}$$

So,  $\hat{\theta}_{MLE} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(\theta|x) = 0.5$ . Furthermore,

$$\begin{aligned}\pi_\Theta(0.2)L(0.2|x) &= 0.10 \cdot (0.2) \cdot (0.8)^3 = 0.001024, \\ \pi_\Theta(0.5)L(0.5|x) &= 0.01 \cdot (0.5) \cdot (0.5)^3 \approx 0.000625, \\ \pi_\Theta(0.7)L(0.7|x) &= 0.89 \cdot (0.7) \cdot (0.3)^3 = 0.16821.\end{aligned}$$

So,  $\hat{\theta}_{MAP} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} \pi_\Theta(\theta)L(\theta|x) = 0.7$ . i.e. with a non-uniform prior distribution of  $\Theta$ , we have

$$\hat{\theta}_{MLE} = 0.5 \neq 0.7 = \hat{\theta}_{MAP}.$$

### 5.1.5 Goodness of fit

In this section, we have discussed various estimation techniques to find an estimate of an unknown parameter that leads to an estimator or sample statistic of the corresponding parameter and runs over all possible realizations/observations of an I.I.D. random sample of size n. Now, we need to find out which is relatively best i.e. how can we determine which estimator is better (rather than the technique)? There are even more different ways to estimate besides MLE/MoM/MAP, and in different scenarios, different techniques may work better. Here, we will consider some properties of estimators that allow us to compare their goodness.

#### Bias

Once, we come up with an estimator of a parameter using some estimation technique, we need to measure the estimator whether or not its expectation equals to the true parameter and the measure is known as bias of the estimator.

**Definition 5.8** (Bias). Consider a random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable X with unknown but parametrized distribution  $p_X(x|\theta)$  or  $f_X(x|\theta)$ . Suppose  $\hat{\Theta}$  be an estimator of the unknown (but fixed) parameter  $\theta$  under a suitable estimation technique. Then, bias of the estimator  $\hat{\Theta}$  for  $\theta$  is defined as

$$Bias(\hat{\Theta}) = E(\hat{\Theta}) - \theta.$$

Now, we have

- a. If  $Bias(\hat{\Theta}) = 0 \equiv E(\hat{\Theta}) = \theta$ , then we say  $\hat{\Theta}$  is an unbiased estimator of  $\theta$ .
- b. If  $Bias(\hat{\Theta}) > 0$ , then  $\hat{\Theta}$  typically overestimates  $\theta$ .
- c. If  $Bias(\hat{\Theta}) < 0$ , then  $\hat{\Theta}$  typically underestimates  $\theta$ .

**Example 5.14** (Sample mean). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow$

n I.I.D. copies of a random variable X from  $Pois(\theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} = \hat{\theta}_{MoM} \implies \hat{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

Question: Prove that the estimator (or the sample mean),  $\hat{\Theta} = \bar{X}$ , is a unbiased estimator of the  $\theta$  (or the unknown mean).

Solution: We have

$$E(\hat{\Theta}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \theta.$$

This makes sense: the average of your samples should be on-target for the true average.

**Example 5.15** (Sample statistic for  $X \sim Unif(0, \theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow$  n I.I.D. copies of a random variable X from  $X \sim Unif(0, \theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = x_{max}, \hat{\theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n x_i = 2\bar{x} \implies \hat{\Theta}_{MLE} = X_{max}, \hat{\Theta}_{MoM} = 2\bar{X}.$$

Question: Sure,  $\hat{\Theta}_{MLE}$  maximizes the likelihood, so in a way  $\hat{\Theta}_{MLE}$  is better than  $\hat{\Theta}_{MoM}$ . But, what are the biases of these estimators? Before doing any computation: do you think  $\hat{\Theta}_{MLE}$  and  $\hat{\Theta}_{MoM}$  are overestimates, underestimates, or unbiased?

Solution: Actually, it can be considered that  $\hat{\Theta}_{MoM}$  is spot-on since the average of the samples should be close to  $\frac{\theta}{2}$  and multiplying by 2 would seem to give the true  $\theta$ . On the other hand,  $\hat{\Theta}_{MLE}$  might be a bit of an underestimate, since we probably wouldn't have  $\theta$  be exactly the largest (maybe a little larger).

a. Computation of  $Bias(\hat{\Theta}_{MLE})$  wrt  $\theta$ : we have

$$\begin{aligned} f_{\hat{\Theta}_{MLE}}(\theta) &= f_{X_{max}}(x) = \frac{d}{dx} (F_{X_{max}}(x)) = \frac{d}{dx} (P(X_{max} \leq x)) \\ &= \frac{d}{dx} (P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x)) = \frac{d}{dx} ((F_X(x))^n) \\ &= n(F_X(x))^{n-1} f_X(x) = n \left( \frac{x}{\theta} \right)^{n-1} \left( \frac{1}{\theta} \right). \end{aligned}$$

Then

$$\begin{aligned} E(\hat{\Theta}_{MLE}) &= E(X_{max}) = \int_0^\theta xn \left( \frac{x}{\theta} \right)^{n-1} \left( \frac{1}{\theta} \right) dx \\ &= \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1} \theta. \end{aligned}$$

So, we have

$$Bias(E(\hat{\Theta}_{MLE})) = E(\hat{\Theta}_{MLE}) - \theta = \frac{n}{n+1} \theta - \theta = \frac{-\theta}{n+1} < 0.$$

This makes sense because if we had 3 sample points from  $Unif(0, 1)$  for example, we would expect them at  $1/4, 2/4, 3/4$ , and so it would be  $n/(n+1)$  as our expected max. Similarly, if we had 4 sample points, then we would expect them at  $1/5, 2/5, 3/5, 4/5$ , and so it would again be  $n/(n+1)$  as our expected max.

b. Computation of  $Bias(\hat{\Theta}_{MoM})$  wrt  $\theta$ : we have

$$E(\hat{\Theta}_{MoM}) = E\left(\frac{2}{n} \sum_{i=1}^n X_i\right) = \frac{2}{n} \sum_{i=1}^n E(X_i) = \frac{2}{n} \cdot \frac{\theta}{2} = \theta.$$

Consider a realization  $\mathbf{x} = (x_1 = 1, x_2 = 9, x_3 = 2)$  of an I.I.D. sample of size 3 from  $X \sim Unif(0, \theta)$ , then we have the following estimates

$$\hat{\theta}_{MLE} = x_{max} = 9, \hat{\theta}_{MoM} = \frac{2}{3} \sum_{i=1}^3 x_i = \frac{2}{3} \cdot (1 + 9 + 2) = 8.$$

However, based on our realization of the sample, the MoM estimator is impossible. If the actual parameter were 8, then that means that the

distribution we pulled the sample from is  $Unif(0, 8)$ , in which case the likelihood that we get a 9 is 0. But we did see a 9 in our sample. So, even though  $\hat{\Theta}_{MoM}$  is unbiased, it still yields an impossible estimate. This just goes to show that finding the right estimator is actually quite tricky.

A good solution would be to de-bias the MLE by scaling it appropriately. If we decided to have a new estimator based on the MLE:

$$\hat{\Theta} = \frac{n+1}{n} \hat{\Theta}_{MLE} = \frac{n+1}{n} X_{max}.$$

Now, we have an unbiased estimator, but now it does not maximize the likelihood any more. Actually, the MLE is what we say to be asymptotically unbiased as follows:

$$Bias(\hat{\Theta}_{MLE}) = E(\hat{\Theta}_{MLE}) - \theta = \frac{n}{n+1}\theta - \theta = \frac{-\theta}{n+1} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

**Example 5.16** (A reciprocal of sample mean). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  from  $Exp(\theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i} = \frac{1}{\bar{x}} = \hat{\theta}_{MoM} \implies \hat{\Theta} = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i} = \frac{1}{\bar{X}}.$$

Question: Prove that the estimator,  $\hat{\Theta} = \frac{1}{\bar{X}}$ , is a biased estimator of the  $\theta$  (or the unknown parameter).

Solution: We have

$$E(\hat{\Theta}) = E\left(\frac{1}{\frac{1}{n} \sum_{i=1}^n X_i}\right) \geq \frac{1}{E\left(\frac{1}{n} \sum_{i=1}^n X_i\right)} = \frac{1}{\bar{\theta}}.$$

The function  $g(x_1, \dots, x_n) = \frac{1}{\sum_{i=1}^n x_i}$  is convex (at least in the positive octant when all  $x_i \geq 0$ ) since  $\frac{1}{x}$  is a convex function. Then, Jensen's inequality says that  $E(g(x_1, \dots, x_n)) \geq g(E(x_1), E(x_2), \dots, E(x_n))$  because of convexity. So  $E(\hat{\Theta}) \geq \theta$  systematically, and we typically have an overestimate.

### Variance and mean squared error

We are often also interested in accuracy of an estimator i.e. how much an estimator varies (we would like it to be unbiased and have small variance to that it is more accurate). There is a metric namely mean squared error that captures this property of estimators and defined as below.

**Definition 5.9** (Mean squared error of an estimator). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  with unknown but parametrized distribution  $p_X(x|\theta)$  or  $f_X(x|\theta)$ . Suppose  $\hat{\Theta}$  be an estimator of the unknown (but fixed) parameter  $\theta$  under a suitable estimation technique. Then, mean squared error of the estimator  $\hat{\Theta}$  for  $\theta$  is defined as

$$MSE(\hat{\Theta}) = E\left(\left(\hat{\Theta} - \theta\right)^2\right) = E\left(\left(\hat{\Theta} - E(\hat{\Theta}) + E(\hat{\Theta}) - \theta\right)^2\right)$$

We have

$$\begin{aligned} MSE(\hat{\Theta}) &= E\left(\left(\hat{\Theta} - \theta\right)^2\right) = E\left(\left(\hat{\Theta} - E(\hat{\Theta}) + E(\hat{\Theta}) - \theta\right)^2\right) \\ &= E\left(\left(\hat{\Theta} - E(\hat{\Theta})\right)^2\right) + \left(E(\hat{\Theta}) - \theta\right)^2 + 2E\left(\hat{\Theta} - E(\hat{\Theta})\right)\left(E(\hat{\Theta}) - \theta\right) \\ &= Var(\hat{\Theta}) + Bias^2(\hat{\Theta}) + 0 = Var(\hat{\Theta}) + Bias^2(\hat{\Theta}). \end{aligned}$$

Variance-Bias trade-off: Usually, we want to minimize MSE, and these two quantities are often inversely related. That is, decreasing one leads to an increase in the other, and finding the balance will minimize the MSE i.e.a small MSE often involves a trade-off between variance and bias. For unbiased estimators,  $MSE(\hat{\Theta}) = Var(\hat{\Theta})$ , so no trade-off can be made..

**Example 5.17** (MSE of sample mean). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  from  $Pois(\theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} = \hat{\theta}_{MoM} \implies \hat{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

Question: Compute the mean squared error of the estimator (or the sample mean),  $\hat{\Theta} = \bar{X}$ , with respect to the unknown parameter  $\theta$  (or the unknown mean).

Solution: We have

$$E(\hat{\Theta}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \theta \implies \text{Bias}(\hat{\Theta}) = 0.$$

And

$$\text{Var}(\hat{\Theta}) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\theta}{n}.$$

So, mean squared error of  $\hat{\Theta}$  is given by

$$\text{MSE}(\hat{\Theta}) = E((\hat{\Theta} - \theta)^2) = \text{Var}(\hat{\Theta}) + \text{Bias}^2(\hat{\Theta}) = \frac{\theta}{n}.$$

### Consistency of an estimator

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter.

**Definition 5.10** (Consistency of an estimator). Consider a random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  with unknown but parametrized distribution  $p_X(x|\theta)$  or  $f_X(x|\theta)$ . Suppose  $\hat{\Theta}$  be an estimator of the unknown (but fixed) parameter  $\theta$  under a suitable estimation technique. Then, the estimator  $\hat{\Theta}$  is consistent if  $\hat{\Theta}$  converges in probability to  $\theta$ . That is, for any  $\varepsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta} - \theta| > \varepsilon) = 0 \equiv \lim_{n \rightarrow \infty} P(|\hat{\Theta} - \theta| \leq \varepsilon) = 1.$$

Also, the estimator  $\hat{\Theta}$  is consistent in mean squared error i.e.  $\hat{\Theta}$  converges in mean squared error to  $\theta$  if

$$\text{MSE}(\hat{\Theta}) = \text{Var}(\hat{\Theta}) + \text{Bias}^2(\hat{\Theta}) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

**Example 5.18** (Sample statistic for  $X \sim \text{Unif}(0, \theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  from

$X \sim Unif(0, \theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = x_{max}, \hat{\theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n x_i = 2\bar{x} \implies \hat{\Theta}_{MLE} = X_{max}, \hat{\Theta}_{MoM} = 2\bar{X}.$$

Question: Prove that  $\hat{\Theta}$  is a consistent estimator of  $\theta$ .

Answer:

- a. For unbiased estimator  $\hat{\Theta}_{MoM}$  and  $\varepsilon > 0$  we have

$$\begin{aligned} P(|\hat{\Theta}_{MoM} - \theta| > \varepsilon) &= P(|\hat{\Theta}_{MoM} - E(\hat{\Theta}_{MoM})| > \varepsilon) \\ &\leq \frac{Var(\hat{\Theta}_{MoM})}{\varepsilon^2} = \frac{Var(2\bar{X})}{\varepsilon^2} \\ &= \frac{4Var(X)}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

- b. For biased estimator  $\hat{\Theta}_{MLE}$ , we cannot use Chebyshev's inequality unfortunately. For computation of  $Bias(\hat{\Theta}_{MLE})$  wrt  $\theta$ : we have

$$\begin{aligned} F_{\hat{\Theta}_{MLE}}(\theta) &= F_{X_{max}}(x) = P(X_{max} \leq x) = P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\ &= (F_X(x))^n = \left(\frac{x}{\theta}\right)^n. \end{aligned}$$

Now, we have

$$\begin{aligned} P(|\hat{\Theta}_{MoM} - \theta| > \varepsilon) &= P(\hat{\Theta}_{MoM} < \theta - \varepsilon) + P(\hat{\Theta}_{MoM} > \theta + \varepsilon) = \\ P(\hat{\Theta}_{MoM} < \theta - \varepsilon) &= \begin{cases} \left(\frac{\theta - \varepsilon}{\theta}\right)^n & \varepsilon < \theta \text{ i.e. } \theta - \varepsilon > 0 \\ 0 & \varepsilon \geq \theta \text{ i.e. } \theta - \varepsilon \leq 0. \end{cases} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

Now we've seen that, even though the MLE and MoM estimators of  $\theta$  for an i.i.d. sample of size n from  $Unif(0, \theta)$  are different, they are both consistent. Furthermore for  $Unif(0, \theta)$

- a.  $\hat{\Theta}_{MoM}$  is an unbiased and consistent estimator of  $\theta$  for an I.I.D. sample of size n from  $Unif(0, \theta)$  meaning it is correct in expectation, also converges to the true parameter (consistent) since the variance goes to 0.

- b. However, if we ignore all the sample points and just take the first one and multiply it by 2,  $\hat{\Theta} = 2X_1$ , it is unbiased (as  $E(\hat{\Theta}) = E(2X_1) = 2E(X_1) = 2 \cdot \frac{\theta}{2} = \theta$ ), but it's not consistent i.e. the estimator doesn't get better and better with more n because we're not using all n sample points. Consistency requires that as we get more samples, we approach the true parameter.
- c. Biased but consistent, on the other hand, we established that the MLE estimator  $\hat{\Theta}_{MLE}$  (with expectation  $E(\hat{\Theta}_{MLE}) = \frac{n}{n+1}\theta \rightarrow \theta$  as  $n \rightarrow \infty$ ) is biased but consistent.
- d.  $\hat{\Theta} = \frac{1}{X^2}$  is neither unbiased nor consistent would just be some random expression.

### Efficiency

Efficiency says that the estimator has as low variance as possible. This property combined with consistency and unbiasedness make our estimator is on target (unbiased).

**Definition 5.11.** Consider a realization  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of an I.I.D. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable X with unknown but parametrized distribution  $p_X(x|\theta)$  or  $f_X(x|\theta)$ . Then, define the following

- a. Score function:

$$s(\mathbf{x}) = \frac{\partial}{\partial \theta} (\ln(f_X(x|\theta))) = \frac{f'_X(x|\theta)}{f_X(x|\theta)}.$$

Then, expectation of score statistic is given by

$$\begin{aligned} E(S(\theta)) &= E\left(\frac{f'_X(x|\theta)}{f_X(x|\theta)}\right) = \int \frac{f'_X(x|\theta)}{f_X(x|\theta)} f_X(x|\theta^*) dx \\ &\approx \frac{\partial}{\partial \theta} \int f_X(x|\theta^*) dx = \frac{\partial}{\partial \theta}(1) = 0. \end{aligned}$$

- b. Fisher information matrix:

$$\begin{aligned} I(\theta) &= E\left((S(\theta) - E(S(\theta)))(S(\theta) - E(S(\theta)))^\top\right) = E\left(S(\theta)S^\top(\theta)\right) = \\ &nE\left(\left(\frac{\partial}{\partial \theta} (\ln(L(\theta|\mathbf{x})))\right)^2\right) = -E\left(\frac{\partial^2}{\partial \theta^2} (\ln(L(\theta|\mathbf{x})))\right) = -E\left(H_{\ln(L(\theta|\mathbf{x}))}\right). \end{aligned}$$

**Definition 5.12 (CRLB).** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n)$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , where  $\theta$  is a parameter (or vector of parameters). If  $\hat{\Theta}$  is an unbiased estimator for  $\theta$ , then

$$MSE(\hat{\Theta}) = Var(\hat{\Theta}) \geq \frac{1}{I(\theta)}.$$

where  $I(\theta)$  is the Fisher information. What this is saying is, for any unbiased estimator  $\hat{\Theta}$  for  $\theta$ , the variance  $Var(\hat{\Theta}) = MSE(\hat{\Theta})$  is at least  $\frac{1}{I(\theta)}$ . If we achieve this lower bound, meaning our variance is exactly equal to  $\frac{1}{I(\theta)}$ , then we have the best variance possible for our estimate. That is, we have the minimum variance unbiased estimator (MVUE) for  $\theta$ .

If  $\hat{\Theta}$  is an unbiased estimator for  $\theta$ , then

$$MSE(\hat{\Theta}) = Var(\hat{\Theta}) + Bias^2(\hat{\Theta}) \geq \frac{1}{I(\theta)}.$$

If  $\hat{\Theta}$  is any unbiased estimator for  $\theta$ , there is a minimum possible variance (variance = MSE for unbiased estimators) and the estimator achieves this lowest possible variance, it is said to be efficient.

**Definition 5.13 (Efficiency).** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n)$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , where  $\theta$  is a parameter (or vector of parameters). If  $\hat{\Theta}$  is an unbiased estimator for  $\theta$ , then its efficiency is defined as

$$e(\hat{\Theta}) = \frac{I^{-1}(\theta)}{Var(\hat{\Theta})} \leq 1.$$

This will always be between 0 and 1 because if our variance is equal to the CRLB, then it equals 1, and anything greater will result in a smaller value. A larger variance will result in a smaller efficiency, and we want our efficiency to be as high as possible. An unbiased estimator is said to be efficient if it achieves the CRLB - meaning  $e(\hat{\Theta}) = 1$ . That is, it could not possibly have a lower variance. Again, the CRLB is not guaranteed for biased estimators.

Let  $\hat{\Theta}_1$  and  $\hat{\Theta}_2$  be two unbiased estimators of an unknown parameter  $\theta$  with equal sample sizes. Then,  $\hat{\Theta}_1$  is a more efficient estimator than  $\hat{\Theta}_2$  if  $Var(\hat{\Theta}_1) < Var(\hat{\Theta}_2)$  i.e.  $e(\hat{\Theta}_1) > e(\hat{\Theta}_2)$ .

**Example 5.19** (Efficiency of sample mean). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable  $X$  from  $Pois(\theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} = \hat{\theta}_{MoM} \implies \hat{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

Question: Prove that the estimator (or the sample mean),  $\hat{\Theta} = \bar{X}$ , with respect to the unknown parameter  $\theta$  (or the unknown mean) is efficient or not.

Solution: First, we need to check that it's unbiased, as the CRLB only holds for unbiased estimators.

$$E(\hat{\Theta}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \theta \implies Bias(\hat{\Theta}) = 0.$$

Now, we compute the variance of  $\hat{\Theta}$ :

$$Var(\hat{\Theta}) = Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\theta}{n}.$$

Then, we need to define likelihood function  $L(\theta|\mathbf{x})$  to compute Fisher Information that gives us the CRLB, and see if our variance matches.

$$\begin{aligned} L(\theta|\mathbf{x}) &= P(\mathbf{x}|\theta) = \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{(x_i)!} \implies \\ \ln(L(\theta|\mathbf{x})) &= \sum_{i=1}^n (-\theta + x_i \ln(\theta) - \ln((x_i)!)) \implies \frac{\partial}{\partial \theta} (\ln(L(\theta|\mathbf{x}))) \\ &= \sum_{i=1}^n \left( -1 + \frac{x_i}{\theta} \right) \implies \frac{\partial^2}{\partial \theta^2} (\ln(L(\theta|\mathbf{x}))) = \sum_{i=1}^n \left( -\frac{x_i}{\theta^2} \right) = \frac{-n\bar{x}}{\theta^2}. \end{aligned}$$

Then, in order to compute information matrix  $I(\theta)$ , we compute the expectation of second derivative of log-likelihood function as

$$E\left(\frac{\partial^2}{\partial \theta^2} (\ln(L(\theta|\mathbf{x})))\right) = E\left(\sum_{i=1}^n -\frac{x_i}{\theta^2}\right) = \frac{-\sum_{i=1}^n E(X_i)}{\theta^2} = \frac{-n\theta}{\theta^2} = \frac{-n}{\theta}.$$

So, Fisher Information is the negative expected value of the second derivative of the log-likelihood, so, we have

$$I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} (\ln(L(\theta|\mathbf{x})))\right) = \frac{n}{\theta}.$$

Finally, efficiency of the estimator  $\hat{\Theta}$  as the inverse of the Fisher Information over the variance computed as

$$e(\hat{\Theta}) = \frac{I^{-1}(\theta)}{Var(\hat{\Theta})} = 1.$$

Thus, we've shown that, since the efficiency is 1, the estimator is efficient. That is, it has the best possible variance among all unbiased estimators of  $\theta$ .

### Sufficiency

Consider a realization  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of an i.i.d. sample of size  $n$  from a known distribution model with unknown parameter  $\theta$ . Imagine we have two people:

- a. Statistician A: Knows the entire sample, gets  $n$  quantities:  $\mathbf{x} = (x_1, \dots, x_n)$ .
- b. Statistician B: Knows  $T(x_1, \dots, x_n) = t$ , a single number which is a function of the realizations of an I.I.D. sample of size  $n$ . For example, the sum or the maximum of the sample points.

Heuristically,  $T(x_1, \dots, x_n)$  is a sufficient statistic if Statistician B can do just as good a job as Statistician A, given information. For example, if the samples are from the Bernoulli distribution, knowing  $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$  (the number of heads) is just as good as knowing all the individual outcomes, since a good estimate would be the number of heads over the number of total trials. Hence, we don't actually care the ORDER of the outcomes, just how many heads occurred. The word sufficient in English roughly means enough, and so this terminology is well-chosen.

**Definition 5.14** (Statistic). Consider a realization  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of an i.i.d. sample of size  $n$  i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of  $X$  having a known distribution model with unknown parameter  $\theta$ . Then, we define a statistic as a function  $T : \mathbb{R}^n \rightarrow \mathbb{R}$  of realizations  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of the I.I.D. sample of size  $n$ . For examples:

- a.  $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i \rightarrow$  the sum.

- b.  $T(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i \rightarrow$  the sample mean.
- c.  $T(x_1, \dots, x_n) = \max_{i=1,2,\dots,n} x_i = x_{\max} \rightarrow$  the max/largest value.
- d.  $T(x_1, \dots, x_n) = x_1 \rightarrow$  just take the first sample point.
- e.  $T(x_1, \dots, x_n) = 7 \rightarrow$  ignore all the sample points.

All estimators are statistics because they take in our  $n$  data points and produce a single number.

**Definition 5.15** (Sufficient statistic). A statistic  $T = T(X_1, \dots, X_n)$  is a sufficient statistic if the conditional distribution of  $(X_1, \dots, X_n)$  given  $T = t$  and  $\theta$  does not depend on  $\theta$  i.e.

$$P(X_1 = x_1, \dots, X_n = x_n | T = t, \theta) = P(X_1 = x_1, \dots, X_n = x_n | T = t).$$

**Theorem 5.1.** Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size  $n$  with likelihood  $L(\theta|\mathbf{x})$ . A statistic  $T = T(x_1, \dots, x_n)$  is sufficient if and only if there exist non-negative functions  $g$  and  $h$  such that:

$$L(\theta|\mathbf{x}) = g(\mathbf{x}) \cdot h(T(\mathbf{x}), \theta)$$

That is, the likelihood of the data can be split into a product of two terms: the first term  $g$  can depend on the entire data, but not  $\theta$ , and the second term  $h$  can depend on  $\theta$ , but only on the data through the sufficient statistic  $T$ . (In other words,  $T$  is the only thing that allows the data  $x_1, \dots, x_n$  and  $\theta$  to interact.) That is, we don't have access to the  $n$  individual quantities  $x_1, \dots, x_n$  just the single number ( $T$ , the sufficient statistic).

But basically, we want to split the likelihood into a product of two terms/functions:

- a. For the first term  $g$ , you are allowed to know each individual sample point if you want, but NOT  $\theta$ .
- b. For the second term  $h$ , you can only know the sufficient statistic (single number)  $T(x_1, \dots, x_n)$  and  $\theta$ . We may not know each individual  $x_i$ .

**Example 5.20** (Sample statistic for  $X \sim Unif(0, \theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable X from  $X \sim Unif(0, \theta)$ , then the MLE and MoM both estimate the same given by

$$\hat{\theta}_{MLE} = x_{max}, \hat{\theta}_{MoM} = \frac{2}{n} \sum_{i=1}^n x_i = 2\bar{x} \implies \hat{\Theta}_{MLE} = X_{max}, \hat{\Theta}_{MoM} = 2\bar{X}.$$

Question: Prove that  $\hat{\Theta}_{MLE} = X_{max}$  is a sufficient estimator of  $\theta$ .

Answer: We have the likelihood function  $L(\theta|\mathbf{x})$  as:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\theta} I_{\{0 \leq x_i \leq \theta\}} = \frac{1}{\theta^n} I_{\{0 \leq x_1, \dots, x_n \leq \theta\}} = \frac{1}{\theta^n} I_{\{0 \leq x_{max} \leq \theta\}} \\ &= \frac{1}{\theta^n} I_{\{T(x_1, \dots, x_n) \leq \theta\}}. \end{aligned}$$

Now, we choose

$$g(x_1, \dots, x_n) = 1, \quad h(T(x_1, \dots, x_n), \theta) = \frac{1}{\theta^n} I_{\{T(x_1, \dots, x_n) \leq \theta\}}$$

By the Neyman-Fisher Factorization Criterion,  $\hat{\Theta}_{MLE} = X_{max}$  is a sufficient estimator. For the h term, notice that we just need to know the max of the samples  $T(x_1, \dots, x_n)$  to compute h: we don't actually need to know each individual  $x_i$ . It is noted that here the only interaction between the data and parameter  $\theta$  happens through the sufficient statistic (the max of all the values).

**Example 5.21** (Sample statistic for  $X \sim Pois(\theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable X from  $X \sim Pois(\theta)$ , then the we have

$$\begin{aligned} T(x_1, \dots, x_n) &= \sum_{i=1}^n x_i, \quad \hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \implies \\ T(X_1, \dots, X_n) &= S_n, \quad \hat{\Theta}_{MLE} = \bar{X}. \end{aligned}$$

Question: Prove that  $T(X_1, \dots, X_n) = S_n$  is a sufficient statistic and hence  $\hat{\Theta}_{MLE} = \bar{X}$  is a sufficient estimator. (The reason this is true is

because we don't need to know each individual sample to have a good estimate for  $\theta$ , we just need to know how many events happened total.)  
 Answer: We have the likelihood function  $L(\theta|\mathbf{x})$  as:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{(x_i)!} = \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i)!} \\ &= \frac{1}{\prod_{i=1}^n (x_i)!} \cdot e^{-n\theta} \theta^{T(x_1, \dots, x_n)}. \end{aligned}$$

Now, we choose

$$g(x_1, \dots, x_n) = \frac{1}{\prod_{i=1}^n (x_i)!}, \quad h(T(x_1, \dots, x_n), \theta) = e^{-n\theta} \theta^{T(x_1, \dots, x_n)}.$$

By the Neyman-Fisher Factorization Criterion,  $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$  is a sufficient statistic. This implies  $\hat{\Theta}_{MLE} = \frac{T(X_1, \dots, X_n)}{n}$  is also a sufficient statistic, since knowing the total number of events and the average number of events is equivalent (since we know n).

It is noted that here we had the g term handle some function of only  $x_1, \dots, x_n$  but not  $\theta$ . For the h term though, we do have  $\theta$  but don't need the individual samples  $x_1, \dots, x_n$  to compute h. Imagine being just given  $T(x_1, \dots, x_n)$ , now you have enough information to compute h.

**Example 5.22** (Sample statistic for  $X \sim Ber(\theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. random sample of size n i.e.  $(X_1, X_2, \dots, X_n) \rightarrow n$  I.I.D. copies of a random variable X from  $X \sim Ber(\theta)$ , then the we have

$$\begin{aligned} T(x_1, \dots, x_n) &= \sum_{i=1}^n x_i, \quad \hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \implies \\ T(X_1, \dots, X_n) &= S_n, \quad \hat{\Theta}_{MLE} = \bar{X}. \end{aligned}$$

Question: Prove that  $T(X_1, \dots, X_n) = S_n$  is a sufficient statistic and hence  $\hat{\Theta}_{MLE} = \bar{X}$  is a sufficient estimator. (The reason this is true is because we don't need to know each individual sample to have a good estimate for  $\theta$ , we just need to know how many events happened total.)  
 Answer: We have the likelihood function  $L(\theta|\mathbf{x})$  as:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n p_X(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{x_1+\dots+x_n} (1-\theta)^{n-x_1-\dots-x_n} \\ &= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

Now, we choose

$$g(x_1, \dots, x_n) = 1, \quad h(T(x_1, \dots, x_n), \theta) = \theta^{T(x_1, \dots, x_n)}(1 - \theta)^{n - T(x_1, \dots, x_n)}.$$

By the Neyman-Fisher Factorization Criterion,  $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$  is a sufficient statistic. This implies  $\hat{\Theta}_{MLE} = \frac{T(X_1, \dots, X_n)}{n}$  is also a sufficient statistic, since knowing the total number of events and the average number of events is equivalent (since we know n).

It is noted that here we had the g term handle some function of only  $x_1, \dots, x_n$  but not  $\theta$ . For the h term though, we do have  $\theta$  but don't need the individual samples  $x_1, \dots, x_n$  to compute h. Imagine being just given  $T(x_1, \dots, x_n)$ , now you have enough information to compute h.

We can notice that here the only interaction between the data and parameter  $\theta$  happens through the sufficient statistic (the sum/mean of all the values). We don't actually need to know each individual  $x_i$ .

## 5.2 Interval estimation

Till now, we have discussed several techniques to estimate unknown parameters and the goodness criterion of the estimator. Even if the estimator is having all the good properties, the probability that the estimator for  $\theta$  is exactly correct is 0 i.e.  $P(\hat{\Theta} = \theta) = 0$ , since  $\theta$  is continuous (a decimal number). So, we need to construct confidence intervals around the estimator, containing the true value of the parameter with high probability i.e.  $\hat{\theta}$  is very close to  $\theta$  with high probability. Confidence intervals are used in the Frequentist setting, which means the population parameters are assumed to be unknown but will always be fixed, not random variables. Credible intervals, on the other hand, are a Bayesian version of a Frequentist's confidence interval.

### 5.2.1 Confidence interval

**Theorem 5.2** (A  $100(1 - \alpha)\%$  confidence interval). Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample of size n from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ . Using CLT, an  $100(1 - \alpha)\%$  confidence inter-

val of an unbiased estimate  $\hat{\theta}$  for  $\theta$  is given by

$$\hat{\theta} \in \left[ \theta - z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\theta})}, \theta + z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\theta})} \right].$$

Where  $z_{1-\frac{\alpha}{2}}$  is z-score given by

$$z_{1-\frac{\alpha}{2}} = \varphi^{-1}(a - \frac{\alpha}{2}) = \begin{cases} 1.96 & \text{for } \alpha = 5\% = 0.05 \\ 2.576 & \text{for } \alpha = 1\% = 0.01. \end{cases}$$

*Proof.* We have the following information:

i.  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ , so

$$E(\Theta) = g_1(\theta) \text{ and } Var(\Theta) = g_2(\theta).$$

ii. a realization  $\mathbf{x} = (x_1, \dots, x_n)$  with given n, computes a value of  $\hat{\theta}$ . Also,  $Var(\Theta) = \frac{g_1(\theta)}{n}$ .

iii. In  $100(1 - \alpha)\%$  the  $\alpha$  is given (e.g.  $\alpha = 5\%$  or  $1\%$  etc).

iv. A technique-central limit theorem.

To compute a  $100(1 - \alpha)\%$  confidence interval, we have

$$\begin{aligned} P(|\hat{\theta} - \theta| > \varepsilon) \leq \alpha &\implies P\left(\left|\frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}}\right| > \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}}\right) \leq \alpha \approx \\ P\left(\left|Z\right| > \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}}\right) \leq \alpha &\implies 1 - P\left(\left|Z\right| \leq \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}}\right) \geq 1 - \alpha \\ P\left(\left|Z\right| \leq \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}}\right) \geq 1 - \alpha &\implies 2\varphi\left(\frac{\varepsilon}{\sqrt{Var(\hat{\theta})}}\right) \geq 2 - \alpha \implies \\ \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}} &\geq \varphi^{-1}\left(1 - \frac{\alpha}{2}\right) \implies \frac{\varepsilon}{\sqrt{Var(\hat{\theta})}} \geq z_{1-\frac{\alpha}{2}} \implies \\ \varepsilon &\geq z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\theta})} \implies \hat{\theta} \in \left[ \theta - z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\theta})}, \theta + z_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{\theta})} \right]. \end{aligned}$$

□

**Example 5.23** (Confidence interval wrt  $Pois(\theta)$ ). Backdrop: Consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an i.i.d. sample of size from  $X \sim Pois(\theta)$  where  $\theta$  is unknown. Then, both MLE and MoM estimates agreed at the sample mean as:

$$\hat{\theta} = \hat{\theta}_{MLE} = \hat{\theta}_{MoM} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Question: Construct an 95% confidence interval of  $\hat{\theta}$  (i.e. it contains  $\theta$  with probability 95%).

Answer: We have the following information:

- i.  $X \sim Pois(\theta)$ , so

$$E(X) = \theta \text{ and } Var(X) = \theta.$$

- ii. a realization  $\mathbf{x} = (x_1, \dots, x_n)$  with given  $n$ , computes a value of  $\hat{\theta} = \bar{x}$ . Also,  $Var(\hat{\Theta}) = \frac{\theta}{n}$ .

- iii. In  $100(1 - \alpha)\%$  the  $\alpha = 5\% = 0.05$ . So, z-score is given by

$$z_{1-\frac{\alpha}{2}} = \varphi^{-1} \left( a - \frac{\alpha}{2} \right) = \varphi^{-1}(0.975) = 1.96.$$

- iv. A technique-central limit theorem.

Now, from theorem (5.2), we have

$$\begin{aligned} \hat{\theta} &\in \left[ \theta - 1.96 \sqrt{\frac{\theta}{n}}, \theta + 1.96 \sqrt{\frac{\theta}{n}} \right] \equiv \theta \in \left[ \hat{\theta} - 1.96 \sqrt{\frac{\hat{\theta}}{n}}, \hat{\theta} + 1.96 \sqrt{\frac{\hat{\theta}}{n}} \right] \equiv \\ \theta &\in \left[ \bar{x} - 1.96 \sqrt{\frac{\bar{x}}{n}}, \bar{x} + 1.96 \sqrt{\frac{\bar{x}}{n}} \right]. \end{aligned}$$

**Example 5.24.** Backdrop: consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. sample of size  $n$  from  $X \sim Ber(\theta)$  with the unknown  $\theta$  as probability of success.

Question: given  $n = 400$  and  $\sum_{i=1}^n x_i = 136$  (observed 136 successes out of 400), construct a 99% confidence interval for  $\theta$  (the unknown probability of success).

Answer: We have the following information:

i.  $X \sim Ber(\theta)$ , so

$$E(X) = \theta \text{ and } Var(X) = \theta(1 - \theta).$$

ii. a realization  $\mathbf{x} = (x_1, \dots, x_n)$  such that  $\sum_{i=1}^n x_i = 136$  with given  $n = 400$ , computes a value of  $\hat{\theta} = \bar{x} = \frac{136}{400} = 0.34$ . Also,  $Var(\hat{\Theta}) = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n}$ .

iii. In  $100(1 - \alpha)\%$  the  $\alpha = 1\% = 0.01$ . So, z-score is given by

$$z_{1-\frac{\alpha}{2}} = \varphi^{-1}(a - \frac{\alpha}{2}) = \varphi^{-1}(0.995) = 2.576.$$

iv. A technique-central limit theorem.

Now, from theorem (5.2), we have

$$\begin{aligned} \hat{\theta} &\in \left[ \theta - 2.576 \sqrt{\frac{\theta}{n}}, \theta + 2.576 \sqrt{\frac{\theta}{n}} \right] \equiv \theta \in \left[ \hat{\theta} - 2.576 \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + 2.576 \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \right] = \\ &= \left[ \bar{x} - 2.576 \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + 2.576 \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right] = \\ &= \left[ 0.34 - 2.576 \cdot \sqrt{\frac{0.34 \cdot 0.66}{400}}, 0.34 + 2.576 \cdot \sqrt{\frac{0.34 \cdot 0.66}{400}} \right] = [0.28, 0.40]. \end{aligned}$$

Approach-02: Central limit theorem and an optimal bound on the variance

$$Var(\hat{\Theta}) = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n}.$$

We have

$$\begin{aligned} 0.005 &\geq P(\hat{\Theta} - \theta > \varepsilon) = \frac{1}{2}P(|\hat{\Theta} - \theta| > \varepsilon) \\ &= \frac{1}{2}P\left(\left|\frac{\hat{\Theta} - \theta}{\sqrt{\theta(1-\theta)/400}}\right| > \frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right) \\ &\approx \frac{1}{2}P\left(|Z| > \frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right) = P\left(Z > \frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right) \\ &= 1 - P\left(Z \leq \frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right) = 1 - \varphi\left(\frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right). \end{aligned}$$

So, we get

$$\begin{aligned} \varphi\left(\frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}}\right) &\geq 1 - 0.005 = 0.995 = \varphi(2.576) = \varphi(z_{0.995}) \implies \\ \frac{\varepsilon}{\sqrt{\theta(1-\theta)/400}} &\geq 2.576 \text{ for any } \theta \in (0, 1) \implies 20\varepsilon \cdot 4 \geq 2.576 \implies \\ \varepsilon &\geq \frac{2.576}{20} \cdot \frac{1}{4} = \frac{2.576}{20} \cdot \frac{1}{4} = 0.032 \implies \text{for a realization } \hat{\theta} = 0.34, \\ \hat{\Theta} &\in [\theta - 0.032, \theta + 0.032] \equiv \theta \in [0.307, 0.372] \approx [0.31, 0.37]. \end{aligned}$$

If we repeat this process several times (getting  $n = 400$  sample points each time and constructing different confidence intervals), about 99% of the confidence intervals we construct will contain  $\theta$ .

**Example 5.25** (Finding  $n$  with a given confidence interval). Backdrop: Assume a realization/people  $\mathbf{x} = (x_1, \dots, x_n)$  (of an I.I.D random sample of size  $n$  from  $X \sim Ber(p)$ ) vote party  $A$  with probability  $p$ . We poll  $n$  people. Let  $M_n$  be the fraction/average of the polled persons for Party  $A$  and defined as

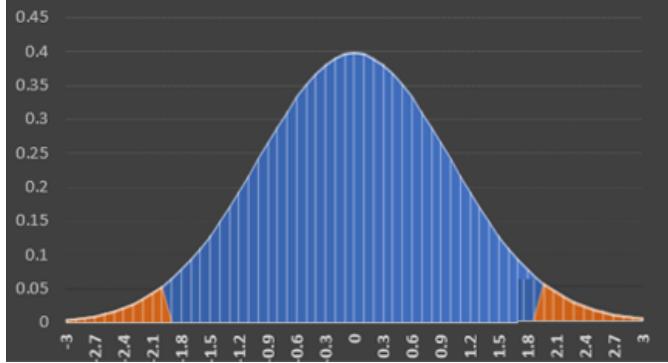
$$M_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Find  $n$  such that estimator  $M_n$  to be within  $\pm 1\%$  of the true value  $p$  with probability at least 95%, statistically

$$P(|M_n - p| > 0.01) \leq 0.05.$$

And geometrically, can be visualized as in the figure (??):

Answer: Since opinion of  $i$ th polled person for party  $A$  is  $X_i \sim Ber(p)$  and the distribution of  $M_n = \frac{1}{n} \sum_{i=1}^n X_i$  is symmetric around its mean



**Figure 5.6:** The distribution  $M_n - p$  is symmetric, and we are looking for tail bounds.

$E(M_n) = p$  with  $Var(M_n) = \frac{p(1-p)}{n}$ , then, we have

$$\begin{aligned}
0.025 &\geq P(M_n - p > 0.01) = \frac{1}{2}P(|M_n - p| > 0.01) \\
&= \frac{1}{2}P\left(\left|\frac{M_n - p}{\sqrt{p(1-p)/n}}\right| > \frac{0.01}{\sqrt{p(1-p)/n}}\right) \\
&\approx \frac{1}{2}P\left(|Z| > \frac{0.01}{\sqrt{p(1-p)/n}}\right) = P\left(Z > \frac{0.01}{\sqrt{p(1-p)/n}}\right) \\
&= 1 - P\left(Z \leq \frac{0.01}{\sqrt{p(1-p)/n}}\right) = 1 - \varphi\left(\frac{0.01}{\sqrt{p(1-p)/n}}\right) \\
&\implies \varphi\left(\frac{0.01}{\sqrt{p(1-p)/n}}\right) \geq 1 - 0.025 = 0.975 = \varphi(1.96) = \varphi(z_{1-0.025}) \\
&\implies \frac{0.01}{\sqrt{p(1-p)/n}} \geq 1.96 \text{ for any } p \in (0, 1) \implies 0.01\sqrt{n} \cdot 4 \geq 1.96 \\
&\implies n \geq \left(\frac{1.96}{0.01} \cdot \frac{1}{4}\right)^2 = 9604.
\end{aligned}$$

In this calculation we used the approximation given by the CLT, but how good is it? If we work out the exact computation, using the fact that  $M_n$  is a scaled binomial and the worse-case parameter  $p = 1/2$ , then we get that  $n = 9604$  will give us a 95.1% probability of being within 0.01 of the true value  $p$ : in this case, the CLT's approximation is very good. So if the CLT is so much better, why use Chebyshev's in-

equality at all? The reason is that Chebyshev is a bound that works for all  $\varepsilon$  and probabilities, is always a valid upper bound but is very conservative. The CLT is often much sharper, however it only works for  $\varepsilon$  of appropriate scale and does not give a rock-solid bound as it is an approximation and not a bound. Is it possible to strengthen Chebyshev's inequality to get a bound nearly as sharp as the CLT's, under appropriate assumptions? Yes, the Chernoff inequality provides the optimal approximation.

### 5.3 Hypothesis testing

**Example 5.26** (A victim is not guilty). Consider a victim is in trial for a crime in the court. Then the Judge will have following hypotheses to determine whether or not the victim is guilty (a given statement happens to be source of an alternative hypothesis) as:

$$\begin{aligned} H_0 &: \text{the victim is not guilty} \\ H_a &: \text{the victim is guilty.} \end{aligned}$$

Based on evidence/observation/realization/data and its statistical analysis, the Judge come up with a conclusion that either reject  $H_0$  or fail to reject  $H_0$ .

**Example 5.27** (Magician's fair coin). An statistical method of hypothesis testing to prove statistically a claim by selecting a suitable hypothesis based on data. Suppose, we have a Magician Sarkar, who says

Magician Sarkar: I have here a fair coin.

And then an audience, an Agnostic (described in (5.7)) statistician named Agyeya, engages him in a conversation:

- Agnostic statistician Agyeya: I don't believe you. Can we examine it?
- Magician Sarkar: Be my guest.
- Agnostic Statistician Agyeya: I am going to flip your coin 100 times and see how many heads I get. (Agyeya flips the coin 100

times and sees 99 heads.) You cannot be telling the truth, there's no way this coin is fair.

— Magician Sarkar: Wait I was just unlucky, I swear I am not lying.

So let's give Sarkar the benefit of the doubt. We compute the probability that we observed an outcome at least as extreme as this, given that Sarkar isn't lying i.e. the coin is fair, so the number of heads observed should be  $X \sim Bin(100, 0.5)$ , because there are 100 independent trials and a 50% of heads since it's fair. So, the probability that we observe at least 99 heads (because we're looking for something as least as extreme), is the sum of the probability of 99 heads and the probability of 100 heads as:

$$\begin{aligned} P(X \geq 99) &= \binom{100}{99}(0.5)^{99}(0.5)^1 + \binom{100}{100}(0.5)^{100}(0.5)^0 \\ &= \frac{101}{2^{100}} \approx 7.96 \times 10^{-29} \approx 0. \end{aligned}$$

Basically, if the coin were fair, the probability of what we just observed (99 heads or more) is basically 0. This is strong statistical evidence that the coin is NOT fair. Our assumption was that the coin is fair, but if this were the case, observing such an extreme outcome would be extremely unlikely. Hence, our assumption is probably wrong. So, this is like a Probabilistic Proof by Contradiction.

**संशय :** [सं० सम्/शी+अच्] मन की वह स्थिति जिसमें किसी बात का ठीक रूप जानने या समझने की उत्कंठा या जिजासा ।

**संशयवाद :** [सं० संशय/वद+घञ] संशय की संदधान्तिक स्थिति (वे ठीक भी हैं अथवा नहीं भी )।

**संशयवादी:** [सं० संशय/वद+णिनि] वह जो संशय वाद का अनुयायी या समर्थक हो।

**Agnostic:** a person who has doubts and asks questions about things that other people believe.

**Figure 5.7:** A self-explanatory meaning of Agnostic in Hindi with corresponding translation in English.

**Example 5.28** (to determine a fair coin). Consider a coin toss. Then, to determine whether the coin is fair or not (happens to be source of an alternative hypothesis), we come up with following hypotheses as:

$$H_0 : \text{the victim is fair i.e. } p = 0.5$$

$$H_a : \text{the victim is not fair i.e. } p \neq 0.5.$$

Based on evidence/observation/realization/data and its statistical analysis, we can conclude that either reject  $H_0$  or fail to reject  $H_0$ .

### A summarized hypothesis testing procedure

Step 1: Make a claim (e.g. like Indian food is spicy, Dirty boots belong on the table etc.)

Step 2: Set up a null hypothesis  $H_0$  and alternative hypothesis  $H_a$  (based on an underlined statistical statement).

- Alternative hypothesis can be one-sided or two-sided.
- The null hypothesis is usually a baseline, no effect, or benefit of the doubt.
- The alternative is what you want to prove, and is opposite the null.

Step 3: Choose a significance level  $\alpha$  (usually  $\alpha = 0.05$  or  $0.01$ ).

Step 4: Collect data (e.g. a realization  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim p_X(x|\theta)$  or  $f_X(x|\theta)$ ). And come up with an estimate  $\hat{\theta}$  for the unknown parameter  $\theta$  (e.g.  $\hat{\theta} = \bar{x}$ ).

Step 5: Compute a p-value as:

$$p = P(\text{observing data at least as extreme as our } |H_0 \text{ is true}).$$

Step 6: State the conclusion as:

- if  $p < \alpha$ , reject the null hypothesis  $H_0$  in favour of the alternative  $H_a$  (i.e. our result is statistically significant in this case).
- if  $p \geq \alpha$ , fail to reject the null hypothesis  $H_0$ .

**Example 5.29.** Backdrop: consider a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of an I.I.D. random sample of size  $n$  from  $X \sim Ber(\theta)$ , where  $\theta$  is the unknown probability of success.

Question: We want to determine whether or not more than  $3/4$  of Indian would vote for A in 2024. In a random poll sampling  $n = 137$  Indian, we collected responses  $\mathbf{x} = (x_1, \dots, x_n)$  (each is 1 or 0, if they would vote for him or not). We observe 131 yes responses:  $\sum_{i=1}^n x_i = 131$ . Perform a hypothesis test and state your conclusion.

Answer:

a. We have

$$X \sim Ber(\theta) \implies E(X) = \theta, \quad Var(X) = \theta(1 - \theta).$$

b. Let  $p$  denote the true proportion of Americans that would vote for A. Then, the null and alternative hypotheses are:

$$H_0 : \text{Exactly } 3/4 \text{ Indian voted for party A i.e. } p = 0.75$$

$$H_a : \text{More than } 3/4 \text{ Indian voted for party A i.e. } p \geq 0.75.$$

c. Let  $X_i \rightarrow$  the response of  $i$ th polled person in the poll sampling.  
So, we have

$$\begin{aligned} \hat{\Theta} = \bar{X} &\implies E(\Theta) = E(\bar{X}) = \theta, \quad Var(\hat{\Theta}) = Var(\bar{X}) = \frac{Var(X)}{n} \\ &= \frac{\theta(1 - \theta)}{n}. \end{aligned}$$

As, we have a realization  $\mathbf{x} = (x_1, x_2, \dots, x_{137})$  (such that  $\sum_{i=1}^{137} x_i = 131$ ) of an I.I.D. random sample of size  $n = 137$  from  $X \sim Ber(x|\theta)$ . So, we get

$$\hat{\theta} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{131}{137} = 0.96.$$

- d. Now, test the hypothesis at the  $\alpha = 0.01$  significance level. Then, under the null hypothesis, we have  $p = 0.75$ , so, the p-value (observing data at least as extreme), is given by

$$\begin{aligned} P(\bar{X} \geq \bar{x}) &= P\left(\bar{X} \geq \frac{131}{137}\right) = P\left(\frac{\bar{X} - 0.75}{\sqrt{(0.37)^2}} \geq \frac{131/137 - 0.75}{\sqrt{(0.37)^2}}\right) \\ &\approx P(Z \geq 5.42643) \approx 0. \end{aligned}$$

- e. With a p-value so close to 0 (and certainly  $< \alpha = 0.01$ , we reject the null hypothesis that (only) 75% of Indian would vote for A. There is strong evidence that this proportion is actually larger.

Again, what we did was: assume  $p = 0.75$  (null hypothesis), then note that the probability of observing data so extreme (in fact very close to 100% of people), was nearly 0. Hence, we reject this null hypothesis because what we observed would've been so unlikely if it were true.

# 6

---

## Statistical modelling

---

### 6.1 Statistical models

It is fortunate, however, that a large number of estimation problems can be represented by a data model in linear framework as a linear model that allows us to easily determine this estimator. Not only is the minimum variance valued (MVU) estimator immediately evident once the linear model has been identified, but in addition, the statistical performance follows naturally. The key, then, to finding the optimal estimator is in structuring the problem in the linear model form to take advantage of its unique properties.

#### Derivation of least mean square estimator

**Theorem 6.1** (Condition expectation as an LMS estimator). The covariance between two random variables has an important bearing on the predictability of  $Y$  based on knowledge of the outcome of  $X$ . So, consider  $Y$  is statistically related to another random variable  $X$  i.e.  $Y = g(X)$  and  $g$  is unknown. Then, least-mean-square estimation of  $Y$

is given by the conditional expectation as

$$\hat{Y} = E(Y|X) = \hat{g}(X).$$

Moreover, estimation error  $Y - E(Y|X)$  and the attribute  $X$  are uncorrelated as:

$$\begin{aligned} E(X(Y - E(Y|X))) &= E(XY) - E(XE(Y|X)) = 0 = E(X)E(Y - E(Y|X)) \\ \implies Cov(X, Y - E(Y|X)) &= 0. \end{aligned}$$

Since, we have

$$\begin{aligned} E(XE(Y|X)) &= E\left(\int_{x \in \Omega_X} xE(Y|X=x)f_X(x)dx\right) \\ &= \int_{y \in \Omega_Y} \int_{x \in \Omega_X} xyf_X(x)f_{Y|X}(y|x)dxdy \\ &= \int_{(x,y) \in \Omega_X \times \Omega_Y} xyf_{X,Y}(x,y)dxdy = E(XY). \end{aligned}$$

*Proof.* For one observation of  $X = x$ , we have

$$\begin{aligned} MSE &= E((Y - c)^2) = E\left((Y - E(Y|X=x) + E(Y|X=x) - c)^2\right) \\ &= E((Y - E(Y|X=x))^2) + (E(Y|X=x) - c)^2. \end{aligned}$$

To find least mean square estimate, we have

$$\begin{aligned} \frac{d}{dc}(MSE) &= \frac{d}{dc}\left((E(Y|X=x) - c)^2\right) = -2(E(Y|X=x) - c) = 0 \\ \implies c &= E(Y|X=x) = \hat{g}(x). \end{aligned}$$

That is, we have a least mean square estimate of  $Y$  as

$$\hat{Y} = \hat{y} = E(Y|X=x) = \hat{g}(x).$$

Then, varies for every possible observation of  $X$ , we have the least mean square estimator of  $Y$  as:

$$\hat{Y} = E(Y|X) = \hat{g}(X) \text{ such that } E\left((Y - g(X))^2\right) \geq E(Y - E(Y|X))^2 \text{ for any } g.$$

Alternatively, consider an arbitrary function  $g$  of  $X$  and define mean

squared error with respect to  $Y$  as

$$\begin{aligned}
 MSE(g(X)) &= E((Y - g(X))^2) = E((Y - E(Y|X) + E(Y|X) - g(X))^2) \\
 &= E((Y - E(Y|X))^2) + E((E(Y|X) - g(X))^2) \\
 &= E(E((Y - E(Y|X))^2 | X)) + E((E(Y|X) - g(X))^2) \\
 &= E(Var(Y|X)) + E((E(Y|X) - g(X))^2) \\
 &\geq E((E(Y|X) - g(X))^2) \text{ for any other function } g \text{ of } X.
 \end{aligned}$$

□

## 6.2 Linear models

Consider  $Y$  is statistically dependent on another random variable (or vector), as a variate  $X$ , in a linear framework as

$$E(Y|X) = \theta_0 + \theta_1 X.$$

**Theorem 6.2** (Computation of parameters for a linear model). Consider  $Y$  is statistically dependent on another random variable (or vector), as a variate  $X$ , in a linear framework as

$$E(Y|X) = \theta_0 + \theta_1 X.$$

Then, the parameters can be computed as

$$\theta_1 = (Var(X))^{-1} Cov(X, Y), \quad \theta_0 = E(Y) - (Var(X))^{-1} Cov(X, Y)E(X).$$

So,

$$\begin{aligned}
 \hat{Y} &= E(Y|X) = E(Y) + (Var(X))^{-1} Cov(X, Y)(X - E(X)) \equiv \\
 Y &= E(Y|X) + \varepsilon = E(Y) + (Var(X))^{-1} Cov(X, Y)(X - E(X)) + \varepsilon.
 \end{aligned}$$

*Proof.* We have

$$\begin{aligned}
 \hat{Y} &= E(Y|X) = aX + b \implies Y = aX + b + \varepsilon \implies E(Y) = aE(X) + b \implies \\
 E(E(X)E(Y|X)) &= a(E(X))^2 + bE(X), \quad E(XY) = aE(X^2) + bE(X) \implies \\
 a &= (E(X^2) - (E(X))^2)^{-1} (E(XY) - E(X)E(Y)), \quad b = E(Y) - aE(X) \\
 \implies \hat{Y} &= E(Y|X) = E(Y) + (Var(X))^{-1} Cov(X, Y)(X - E(X)).
 \end{aligned}$$

□

### 6.2.1 Simple univariate linear model (SULM)

A linear model for regression assumes that the LMS estimate of  $Y = f(X)$  as a conditional expectation/regression function  $E(Y|X)$  is linear in the inputs  $X_1, \dots, X_d$ . Linear models were largely developed in the pre-computer age of statistics, but even in today's computer era there are still good reasons to study and use them. They are simple and often provide an adequate and interpretable description of how the inputs affect the output. For prediction purposes they can sometimes outperform fancier non-linear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data. Finally, linear methods can be applied to transformations of the inputs to higher dimensional embedded spaces using suitable basis-function approaches.

#### Computation of parameters for SULM

**Theorem 6.3** (Univariate linear model). Consider  $Y$  is statistically dependent on another random variable  $X$  in linear framework as

$$E(Y|X) = \theta_0 + \theta_1 X.$$

Then, the parameters can be computed as

$$\theta_1 = \frac{Cov(X, Y)}{Var(X)}, \quad \theta_0 = E(Y) - \frac{Cov(X, Y)}{Var(X)} E(X).$$

So,

$$\begin{aligned} \hat{Y} &= E(Y|X) = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E(X)) \equiv \\ Y &= E(Y|X) + \varepsilon = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E(X)) + \varepsilon. \end{aligned}$$

*Proof.* We have

$$\begin{aligned} \hat{Y} &= E(Y|X) = aX + b \implies Y = aX + b + \varepsilon \implies E(Y) = aE(X) + b \implies \\ E(E(X)E(Y|X)) &= a(E(X))^2 + bE(X), \quad E(XY) = aE(X^2) + bE(X) \implies \\ a &= \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{Cov(X, Y)}{Var(X)}, \quad b = E(Y) - \frac{Cov(X, Y)}{Var(X)}E(X) \\ \implies \hat{Y} &= E(Y|X) = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E(X)). \end{aligned}$$

□

**Example 6.1** (Signal detection). Question: Consider a noisy observation  $Y = S + E$  of a signal  $S$ , where the signal and noisy components are uncorrelated i.e.  $Cov(S, E) = 0$ . Let  $Var(S) = \sigma_S^2$  and  $Var(E) = \sigma_E^2$ . Then, find an LMS estimate of the signal.

Answer: Using the above derivation for linear model, we get the conditional expectation of the signal  $S$  conditional upon the observation  $Y$  as

$$\begin{aligned} E(S|Y) &= E(S) + \frac{Cov(S, Y)}{Var(Y)}(Y - E(Y)) \\ &= E(S) + \frac{\sigma_S^2}{\sigma_S^2 + \sigma_E^2}(Y - E(Y)). \end{aligned}$$

**Example 6.2** ([Bertsekas and Tsitsiklis, 2008]). Question: Let  $\Theta \sim \mathcal{U}(4, 10)$  and we observe  $\Theta$  with some noise  $W \sim \mathcal{U}(-1, 1)$  independent of  $\Theta$ . That is we observe the experimental value of the random variable  $Y = \Theta + W$ . Find the least square estimate of  $\Theta$ .

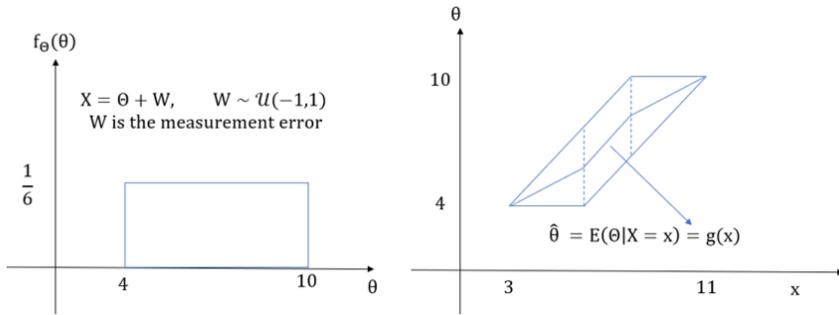
Answer: As,  $\Theta \sim \mathcal{U}(4, 10)$ , noise  $W \sim \mathcal{U}(-1, 1)$  are both are independent. So, conditioned on  $\Theta = \theta$ ,  $X$  is the same as  $\theta + W$ , and is uniformly distributed over the interval  $[\theta - 1, \theta + 1]$ . Thus, joint PDF is given by

$$f_{\Theta, X}(\theta, x) = f_\Theta(\theta)f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{6} \cdot \frac{1}{2} & \theta \in [4, 6], x \in [\theta - 1, \theta + 1] \\ 0, & \text{otherwise.} \end{cases}$$

The slanted rectangle in the right-hand side of figure (6.1) is the set of ordered-pairs  $(\theta, x)$  for which  $f_{\Theta, X}(\theta, x)$  is non-zero. The optimal estimate  $E(\Theta|X = x)$  is the midpoint of that section. It depends on the observed value  $x$  and hence (a piecewise linear) function of  $x$ .

### Estimation of parameters for SULM

**Theorem 6.4** (Least squared estimate of parameter of a SULM). Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an i.i.d. random sample



**Figure 6.1:** Uniform distribution in left plot and least mean squared estimate  $E(\theta|X=x)$  of  $\Theta$  in the right plot.

of size  $n$  i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a joint random variable  $Z = (X, Y)$  with unknown linear model  $\hat{Y} = E(Y|X) = \theta_0 + \theta_1 X \equiv Y = E(Y|X) + \varepsilon = \theta_0 + \theta_1 X$ . Now, running the linear model over  $n$  training examples namely in the data-set  $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^n$  as:

$$\begin{aligned} y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \text{ for } i = 1, 2, \dots, n &\implies \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \theta_0 + \theta_1 x_1 \\ \theta_0 + \theta_1 x_2 \\ \vdots \\ \theta_0 + \theta_1 x_n \end{pmatrix} \\ &+ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies \\ \mathbf{y} = \mathbf{X}\theta + \varepsilon, \text{ where } \mathbf{x}, \mathbf{y}, \varepsilon, \mathbf{1} \in \mathbb{R}^n, \mathbf{X} = (\mathbf{1}, \mathbf{x}) \in \mathbb{R}^{n \times 2}, \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \in \mathbb{R}^2. \end{aligned}$$

So, running the linear model over  $n$  training examples  $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^n$  transform to a linear inverse problem (of the linear model) as:  $\mathbf{y} = \mathbf{X}\theta + \varepsilon$  (a general form of the linear model), where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\theta$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.

c.  $\mathbf{X}$  is called the design matrix.

d. A least square estimate of  $\theta$ :

$$\hat{\theta} = \mathbf{X}^\dagger \mathbf{y} \equiv \arg \min_{\theta} (\|\mathbf{y} - \mathbf{X}\theta\|^2) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix}.$$

And, the MoM estimate of the parameters are given by

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

Finally, we get the best linear predictor of  $\mathbf{Y}$  as:

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X.$$

**Example 6.3.** Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a joint random variable  $Z = (X, Y)$  with  $Y = E(Y|X) + \varepsilon = \theta_0 + \theta_1 X + \varepsilon$ . Question: Students in a probability course claimed that doing the assignments had not helped prepare them for the end-term exam. The end-term exam score  $y$  and assignment score  $x$  (averaged up to the time of the end-term) for the  $n = 18$  students in the class are in the table below in table (6.1):

**Table 6.1:** The exam score  $y$  and the homework score  $x$  averaged upto the time of midterm for 18 students.

x	96	77	0	0	78	64	89	47	90	93	18	86	0	30	59	77	74	67
y	95	80	0	0	79	77	72	66	98	90	0	95	35	50	72	55	75	66

Answer: consider a linear model  $\hat{Y} = E(Y|X) = \theta_0 + \theta_1 X \equiv Y = E(Y|X) + \varepsilon = \theta_0 + \theta_1 X$ , and running over  $n = 18$  training examples

namely in the data-set  $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{18}$  as

$$\begin{aligned} y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \text{ for } i = 1, 2, \dots, 18 &\implies \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \theta_0 + \theta_1 x_1 \\ \theta_0 + \theta_1 x_2 \\ \vdots + \vdots \\ \theta_0 + \theta_1 x_n \end{pmatrix} \\ &+ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies \begin{pmatrix} 95 \\ 80 \\ \vdots \\ 66 \end{pmatrix} = \begin{pmatrix} 1 & 96 \\ 1 & 77 \\ \vdots & \vdots \\ 1 & 67 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies \\ \mathbf{y} = \mathbf{X}\theta + \varepsilon, \text{ where } \mathbf{x}, \mathbf{y}, \varepsilon, \mathbf{1} \in \mathbb{R}^{18}, \mathbf{X} = (\mathbf{1}, \mathbf{x}) \in \mathbb{R}^{18 \times 2}, \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \in \mathbb{R}^2. \end{aligned}$$

Furthermore, we have

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} &= \begin{pmatrix} 18 & 1045 \\ 1045 & 80199 \end{pmatrix} \implies (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 0.228 & -0.003 \\ -0.003 & 0.00005 \end{pmatrix}, \\ \mathbf{X}^\top \mathbf{y} &= \begin{pmatrix} 1105 \\ 8195 \end{pmatrix} \implies \hat{\theta} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 10.75 \\ 0.8726 \end{pmatrix}. \end{aligned}$$

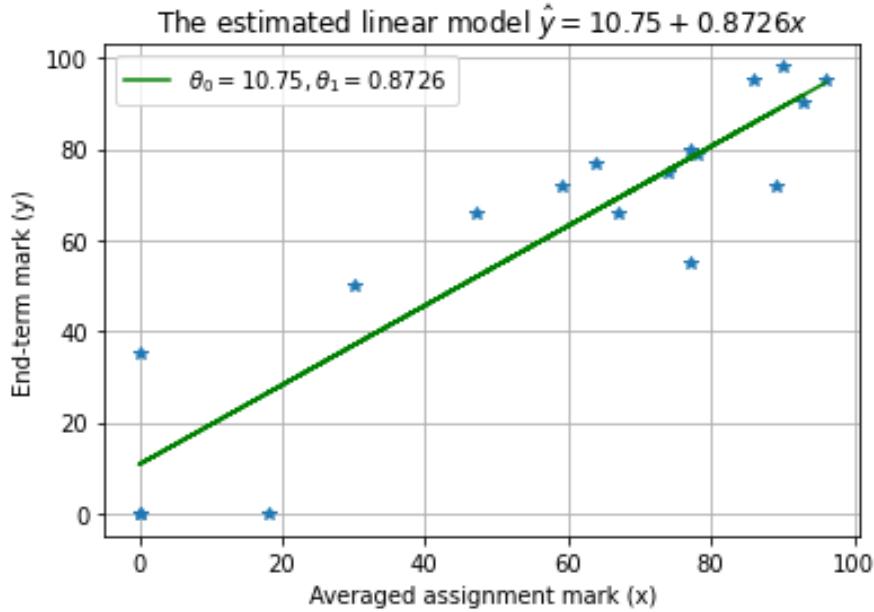
So,  $\mathbf{y} = \mathbf{X}\theta + \varepsilon$  is the general form of a linear model, where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\theta$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.
- c.  $\mathbf{X}$  is called the design matrix.
- d. A least square estimate of  $\theta$ :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\hat{\theta}_0, \hat{\theta}_1)^\top = (10.73, 0.8726)^\top.$$

Alternatively, using method of moments estimate of the parameters, we get

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \approx 0.8726, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \approx 10.73.$$



**Figure 6.2:** The estimated linear model  $\hat{y} = 10.73 + 0.8726x$  that can predict the end-term exam mark for a student with a given averaged assignment mark.

**Theorem 6.5.** Least squared estimate of parameters of a simple univariate linear models, under the Gaussian noise, is the maximum likelihood estimate of the parameters of the SULM.

**Theorem 6.6.** Method of moments estimate of parameters of a simple univariate linear model is given by

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

**Theorem 6.7.** Max a posteriori estimate of parameters of a simple univariate linear model.

#### Goodness of fit for estimated parameters of SULM

**Definition 6.1** (Goodness of fit). Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an i.i.d. random sample of size n i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies

of a joint random variable  $Z = (X, Y)$  with  $Y = E(Y|X) + \varepsilon = \theta_0 + \theta_1 X + \varepsilon$ . Then, goodness of fit,  $R^2$  is defined as

$$R^2 = 1 - \frac{\text{Current deviance}}{\text{Total deviance}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

**Theorem 6.8** (Gauss-Markov Theorem). Consider  $\mathbf{y}$  be a random vector with

$$\mathbf{y} = \mathbf{X}\theta + \varepsilon, E(\mathbf{y}) = \mathbf{X}\theta, Var(\mathbf{y}) = \sigma^2 I, \mathbf{X} \in \mathbb{R}^{n \times (d+1)} \text{ rank } (d+1).$$

Then,  $\mathbf{a}^\top \hat{\Theta}$  is the unique linear unbiased estimator of  $\mathbf{a}^\top \theta$  with minimum variance i.e. BLUE estimator i.e.  $\hat{\Theta} \sim \mathcal{N}\left(\theta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$  and  $\sigma^2$  is having an unbiased estimate as:

$$s^2 = \frac{1}{n-2} (\mathbf{y} - \mathbf{X}\hat{\theta})^\top (\mathbf{y} - \mathbf{X}\hat{\theta}) = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}.$$

Also,

a.  $\hat{\Theta}_1 \sim \mathcal{N}\left(\theta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$ .

b.  $\frac{(n-2)S^2}{\sigma^2} \sim \Xi^2(n-2)$ .

c.  $\hat{\Theta}_1$  and  $S^2$  are independent, and hence

$$T = \frac{\hat{\Theta}_1}{\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

is a  $t$  distribution  $T(n-2, \delta)$  with a non-centrality parameter  $\delta = \frac{\theta_1}{\frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}}$ . A  $100(1-\alpha)\%$  confidence interval for  $\theta_1$  is given by

$$\hat{\theta}_1 \pm t_{(\alpha/2, n-2)} \cdot \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Here, with a realization  $\hat{\theta}_1 = 0.8726$ , and  $s = 13.8547$ , we have

$$t = \frac{\hat{\theta}_1}{\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} = \frac{0.8726}{\frac{13.8547}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} = 8.8025 > t_{(0.025, 16)} = 2.210.$$

So, we reject  $H_0 : \theta_1 = 0$  at the a  $\alpha = 0.05$  level of significance. And hence  $\theta_1 \in [0.8726 - 2.120 \cdot 0.09914, 0.8726 + 2.210 \cdot 0.09914] = [0.6624, 1.0828]$ .

### Sufficient statistics for the SULM

#### A linear model system: a strategy of data-driven linear modelling and application

Step 1: Hypothesis class and model structure: propose a model for the data (i.e. a parametric formula linking the response variable with the input variables, recognising the stochastic nature of the response).

Step 2: Estimation of model: fit the model (e.g. find the best set of parameters).

Step 3: Goodness of fit: ask is the model adequate? (i.e. consistent with the data). Does it allow the main questions of the analysis to be answered?

Step 4: Model evaluation and performance: fit other plausible models, compare them, and choose the best.

Step 5: Deployment of the best linear model and decision making.

#### 6.2.2 Simple multivariate linear model

Motivation for simple multivariate linear model:

### Computation of parameters for SUMLM

**Definition 6.2** (Multivariate linear model). Consider  $Y$  is statistically dependence on a number of covariate random variables  $X_1, X_2, \dots, X_d$  (i.e. a random vector  $\mathbf{X} = (X_1, \dots, X_d)^\top \in \mathbb{R}^{d \times 1}$ ) in linear framework as:

$$E(Y|X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \dots + \theta_d X_d = \theta_0 + \theta^\top \mathbf{X}.$$

Then, the parameters can be computed as

$$\begin{aligned}\theta &= (\theta_1, \dots, \theta_d)^\top = Cov(\mathbf{X}, Y)(Cov(\mathbf{X}))^{-1}, \\ \theta_0 &= E(Y) - Cov(\mathbf{X}, Y)(Var(\mathbf{X}))^{-1}E(\mathbf{X}).\end{aligned}$$

So,

$$\begin{aligned}\hat{Y} &= E(Y|\mathbf{X}) = E(Y) + Cov(\mathbf{X}, Y)(Var(\mathbf{X})^{-1}(\mathbf{X} - E(\mathbf{X}))) \equiv \\ Y &= E(Y|\mathbf{X}) + \varepsilon = E(Y) + Cov(\mathbf{X}, Y)(Var(\mathbf{X})^{-1}(\mathbf{X} - E(\mathbf{X}))) + \varepsilon.\end{aligned}$$

**Example 6.4.** xxxxxxxxxxxxxxxxxxxxxxx

### Estimation of parameters for SMLM

**Theorem 6.9** (Parameter estimation of a simple multivariate linear model). Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an i.i.d. random sample of size n i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a joint random variable  $Z = (\mathbf{X}, Y)$  with unknown linear model  $\hat{Y} = E(Y|\mathbf{X}) = \theta_0 + \theta_1 X_1 + \dots + \theta_d X_d \equiv Y = E(Y|\mathbf{X}) + \varepsilon = \theta_0 + \theta_1 X_1 + \dots + \theta_d X_d$ . Now, running the linear model over n training examples namely in the data-set  $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^n$  as

$$y_i = \theta_0 + \theta_1 x_{i1} + \dots + \theta_d x_{id} + \varepsilon_i \text{ for } i = 1, 2, \dots, n \implies$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \theta_0 + \theta_1 x_{11} + \dots + \theta_d x_{1d} \\ \theta_0 + \theta_1 x_{21} + \dots + \theta_d x_{2d} \\ \vdots + \vdots + \dots + \vdots \\ \theta_0 + \theta_1 x_{n1} + \dots + \theta_d x_{nd} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \implies \mathbf{y} = \mathbf{X}_{n \times (d+1)} \boldsymbol{\theta} + \varepsilon,$$

where  $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^\top, j = 1, 2, \dots, d, \mathbf{y}, \varepsilon, \mathbf{1} \in \mathbb{R}^{n \times 1}, \mathbf{X}_{n \times (d+1)} = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{R}^{n \times (d+1)}, \boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_d)^\top \in \mathbb{R}^{(d+1) \times 1}$ .

So,  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon$  is the general form of a linear model, where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.
- c.  $\mathbf{X}$  is called the design matrix.

d. A least square estimate of  $\omega$ :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_d)^\top.$$

**Example 6.5.** Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an i.i.d. random sample of size  $n$  i.e.  $(Z_1, \dots, Z_n) \sim n$  I.I.D. copies of a joint random variable  $Z = (\mathbf{X}, Y)$  with  $Y = E(Y|\mathbf{X}) + \varepsilon = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \varepsilon$ .

Question: Consider a response  $\mathbf{y}$  with bivariate input  $(x_1, x_2)$  and  $n = 12$  from [Freund and Minton, 1979] in the table below:

$x_1$	0	2	2	2	4	4	4	6	6	6	8	8
$x_2$	2	6	7	5	9	8	7	10	11	9	15	13
$y$	2	3	2	7	6	8	10	7	8	12	11	14

Answer: consider a linear model  $\hat{Y} = E(Y|\mathbf{X}) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 \equiv Y = E(Y|\mathbf{X}) + \varepsilon = \theta_0 + \theta_1 X_1 + \theta_2 X_2$ , and running over  $n$  training examples namely in the data-set  $\mathcal{S} = \{(x_{i1}, x_{i2}), y_i\}_{i=1}^n$  as

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \varepsilon_i \text{ for } i = 1, 2, \dots, n \implies \mathbf{y} = \mathbf{X}\theta + \varepsilon,$$

where  $\mathbf{x}, \mathbf{y}, \varepsilon, \mathbf{1} \in \mathbb{R}^{n \times 1}$ ,  $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{n \times 3}$ ,  $\theta = (\theta_0, \theta_1, \theta_2)^\top \in \mathbb{R}^{3 \times 1}$ .

So,  $\mathbf{y} = \mathbf{X}\theta + \varepsilon$  is the general form of a linear model, where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\theta$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.
- c.  $\mathbf{X}$  is called the design matrix.
- d. A least square estimate of  $\theta$ :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\hat{\theta}_0, \hat{\theta}_1)^\top = (10.73, 0.8726)^\top.$$

Now, with respect to above data, we have the following response data

vector, design matrix and the associated computations:

$$y = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 & 4 & 4 & 4 & 6 & 6 & 8 & 8 \\ 2 & 6 & 7 & 5 & 9 & 8 & 7 & 10 & 11 & 9 & 15 & 13 \end{pmatrix}^\top$$

$$X^\top X = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix}$$

$$(X^\top X)^{-1} = \begin{pmatrix} 0.97476 & 0.2429 & -0.22871 \\ 0.2429 & 0.16207 & -0.11120 \\ -0.22871 & -0.11120 & 0.08360 \end{pmatrix}$$

$$\hat{\theta} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}.$$

So, the covariance matrix of  $\hat{\theta}$  is given by

$$Cov(\hat{\theta}) = \sigma^2 (X^\top X)^{-1} = \sigma^2 \begin{pmatrix} 0.975 & 0.243 & -0.229 \\ 0.243 & 0.162 & -0.111 \\ -0.229 & -0.111 & 0.084 \end{pmatrix}$$

The negative value of  $Cov(\hat{\theta}_1, \hat{\theta}_2) = -0.111$  indicates that in repeated sampling (using the same 12 values of  $x_1$  and  $x_2$ ),  $(\hat{\theta}_1$  and  $\hat{\theta}_2)$  would tend to move in opposite directions; that is, an increase in one would be accompanied by a decrease in the other.

### Goodness of fit for estimated parameters of SMLM

**Theorem 6.10** (Gauss-Markov Theorem). Consider  $\mathbf{y}$  be a random vector with

$$\mathbf{y} = \mathbf{X}\theta + \varepsilon, E(\mathbf{y}) = \mathbf{X}\theta, Var(\mathbf{y}) = \sigma^2 I, \mathbf{X} \in \mathbb{R}^{n \times (d+1)} \text{ rank } (d+1).$$

Then,  $\mathbf{a}^\top \hat{\Theta}$  is the unique linear unbiased estimator of  $\mathbf{a}^\top \theta$  with minimum variance i.e. BLUE estimator i.e.  $\hat{\Theta} \sim \mathcal{N}(\theta, \sigma^2 (X^\top X)^{-1})$  and  $\sigma^2$  is having an unbiased estimate as:

$$s^2 = \frac{1}{n-d-1} (\mathbf{y} - \mathbf{X}\hat{\theta})^\top (\mathbf{y} - \mathbf{X}\hat{\theta}) = \frac{1}{n-d-1} \sum_{i=1}^n (y_i - \hat{y})^2 = \frac{SSE}{n-2}.$$

Also,

- a.  $\hat{\Theta} \sim \mathcal{N}\left(\theta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$ .
- b.  $\frac{(n-d-1)S^2}{\sigma^2} \sim \Xi^2(n - d - 1)$ .
- c.  $\hat{\Theta}_1$  and  $S^2$  are independent. Here, with a realization  $\hat{\theta} = (5.3754, 3.0118, -1.2855)^\top$ , we get  $s^2 = 2.829$ .

### 6.2.3 Embedded linear models (ELM)

An embedded simple multivariate linear model is deep-down inside a simple multivariate linear model in the transformed higher dimensional embedded space, so computation and estimation of parameters for the ESMLM are similar to that of the SMLM.

#### Computation of parameters for ELM

**Definition 6.3** (Embedded simple multivariate linear model). Consider  $Y$  is statistically dependent on a number of covariate random variables  $X_1, X_2, \dots, X_d$  (i.e. a random vector  $\mathbf{X} = (X_1, \dots, X_d)^\top \in \mathbb{R}^{d \times 1}$ ) (that got transformed to  $\varphi(\mathbf{X}) = (\varphi_1(\mathbf{X}), \dots, \varphi_p(\mathbf{X}))^\top$  in the  $p$ -dimensional embedded space using the non-linear basis functions  $\varphi_i$ ) in linear framework as:

$$\begin{aligned} E(Y|\mathbf{X}) &= \theta_0 + \theta_1 \varphi_1(\mathbf{X}) + \theta_2 \varphi_2(\mathbf{X}) + \dots + \theta_p \varphi_p(\mathbf{X}) \\ &= \theta_0 + \theta^\top \varphi(\mathbf{X}). \end{aligned}$$

Then, the parameters can be computed as

$$\begin{aligned} \theta &= (\theta_1, \dots, \theta_p)^\top = Cov(\varphi(\mathbf{X}), Y)(Cov(\varphi(\mathbf{X})))^{-1}, \\ \theta_0 &= E(Y) - Cov(\varphi(\mathbf{X}), Y)(Var(\varphi(\mathbf{X})))^{-1}E(\varphi(\mathbf{X})). \end{aligned}$$

So,

$$\begin{aligned} \hat{Y} &= E(Y|\mathbf{X}) = E(Y) + Cov(\mathbf{X}, Y)(Var(\mathbf{X})^{-1}(\mathbf{X} - E(\mathbf{X}))) \equiv \\ Y &= E(Y|\mathbf{X}) + \varepsilon = E(Y) + Cov(\mathbf{X}, Y)(Var(\mathbf{X})^{-1}(\mathbf{X} - E(\mathbf{X}))) + \varepsilon. \end{aligned}$$

**Theorem 6.11** (A Kernel embedded and ridge regularized linear model (KRLM)). Let us now consider a reproducing Kernel Hilbert space (RKHS),  $\mathcal{H}$ , associated to a positive definite Kernel  $K$  on input space  $\mathcal{X}$ . Then, an estimate of KRLM is obtained by regularizing the mean squared error (MSE) criterion by the RKHS (a Hilbert space produced by a Kernel inspired feature map  $\varphi(x) = k(\cdot, x)$ ) norm:

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

*Proof.* Now, the modified estimation problem can be posed as:

$$\begin{aligned} \hat{f} &= \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right) \\ &= \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \end{aligned}$$

By the representation theorem, the solution of the unconstrained problem can be expanded as:

$$\hat{f}(x) = \sum_{i=1}^n \alpha_i k(x_i, x) = \hat{\omega}^\top \varphi(x), \varphi : \mathcal{X} \rightarrow \mathbb{R} \text{ such that } \varphi(x) = k(\cdot, x).$$

To estimate  $\omega$ , we need to consider the following:

$$\begin{aligned} y &= (y_1, y_2, \dots, y_n)^\top, K = (k(x_i, x_j))_{n \times n}, k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j), \\ \alpha &= (\alpha_1, \alpha_2, \dots, \alpha_n)^\top, (\hat{f}(x_1), \dots, \hat{f}(x_n))^\top \approx K\alpha. \end{aligned}$$

Then, the estimation problem can be reposed as

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{R}^n} \left( \frac{1}{n} (K\alpha - y)^\top (K\alpha - y) + \lambda \alpha^\top K \alpha \right)$$

So, the estimate of  $\alpha$  is computed as:

$$\begin{aligned} \nabla J(\alpha) &= \frac{2}{n} K(K\alpha - y) + 2\lambda K\alpha = \frac{2}{n} K((K\alpha + \lambda nI) - y) = 0 \implies \\ \alpha - (K + \lambda nI)^{-1} y &\in \text{Ker}(K) \implies \alpha = (K + \lambda nI)^{-1} y + \varepsilon, \varepsilon \in \text{Ker}(K) \implies \\ \hat{\alpha} &= (K + \lambda nI)^{-1} y \text{ with } K\varepsilon = 0. \end{aligned}$$

□

### Estimation of parameters for ELM

**Theorem 6.12** (Parameter estimation of an embedded simple multivariate linear model). Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an I.I.D. random sample of size  $n$  i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a joint random variable  $Z = (\mathbf{X}, Y) \rightarrow (\varphi(\mathbf{X}), Y)$  with unknown linear model  $\hat{Y} = E(Y|\varphi(\mathbf{X})) = \theta_0 + \theta_1\varphi_1(\mathbf{X}) + \dots + \theta_p\varphi_p(\mathbf{X}) \equiv Y = E(Y|\varphi(\mathbf{X})) + \varepsilon = \theta_0 + \theta_1\varphi_1(\mathbf{X}) + \dots + \theta_p\varphi_p(\mathbf{X}) + \varepsilon$ . Now, running the linear model over  $n$  training examples namely in the data-set  $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^n$  as

$$y_i = \theta_0 + \theta^\top \varphi \mathbf{x}_i + \varepsilon_i \text{ for } i = 1, 2, \dots, n \implies \mathbf{y} = \mathbf{X}_{n \times (p+1)} \theta + \varepsilon, \\ \text{where } \varphi(\mathbf{x})_j = (\varphi_{1j}\mathbf{x}, \dots, \varphi_{nj}(\mathbf{x}))^\top, j = 1, 2, \dots, p, \mathbf{y}, \varepsilon, \mathbf{1} \in \mathbb{R}^{n \times 1}, \mathbf{X}_{n \times (p+1)} = \\ (\mathbf{1}, \varphi(\mathbf{x})_1, \dots, \varphi(\mathbf{x})_p) \in \mathbb{R}^{n \times (p+1)}, \theta = (\theta_0, \theta_1, \dots, \theta_p)^\top \in \mathbb{R}^{(p+1) \times 1}.$$

So,  $\mathbf{y} = \mathbf{X}\theta + \varepsilon$  is the general form of a linear model, where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\theta$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.
- c.  $\mathbf{X}$  is called the design matrix.
- d. A least square estimate of  $\omega$ :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_p)^\top.$$

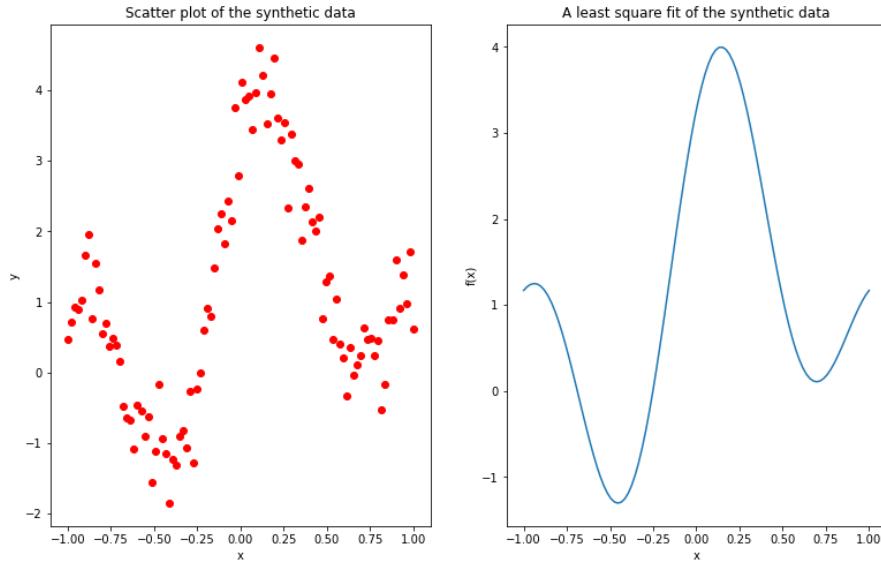
**Example 6.6.** Consider a synthetic data generated from the function:

$$f(x) = 1 + \cos(\pi x) + \sin(\pi x) + \cos(2\pi x) + \sin(2 * \pi x) + 0.5\varepsilon, \varepsilon \sim \mathcal{N}(0, 1).$$

Then, the estimate of the embedded linear model is given by

$$\hat{y} = 1.07 + 1.05 \cos(\pi x) + 1.16 \sin(\pi x) + 1.15 \cos(2\pi x) + 0.98 \sin(2\pi x).$$

**Example 6.7.** Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an I.I.D. random sample of size  $n$  i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a



**Figure 6.3:** In left, there is a scatter plot of the synthetic data and in right, an embedded linear model with sinusoidal basis functions from the synthetic data.

joint random variable  $Z = (X, Y)$  with  $Y = E(Y|X) + \varepsilon = \theta_0 + \theta_1 X + \theta_2 X^2 + \varepsilon$  or  $\theta_0 + \theta_1 X + \theta_2 X^2 + \theta_3 X^3 + \theta_4 X^4 + \varepsilon$ .

Question: Employees in a company claimed that getting a higher position had not helped prepare them to get higher salary. The salary  $y$  and the position for the  $n = 10$  employees are given in table (6.2):

**Table 6.2:** The salary  $y$  and the position level  $x$  for 10 employees.

x	1	2	3	4	5	6	7	8	9	10
y	45000	50000	60000	80000	110000	150000	200000	300000	500000	1000000

Answer: consider a linear model  $\hat{Y} = E(Y|\mathbf{X}) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 \equiv Y = E(Y|\mathbf{X}) + \varepsilon = \theta_0 + \theta_1 X_1 + \theta_2 X_2$ , and running over  $n$  training examples namely in the data-set  $\mathcal{S} = \{(x_{i1}, x_{i2}), y_i\}_{i=1}^n$  as

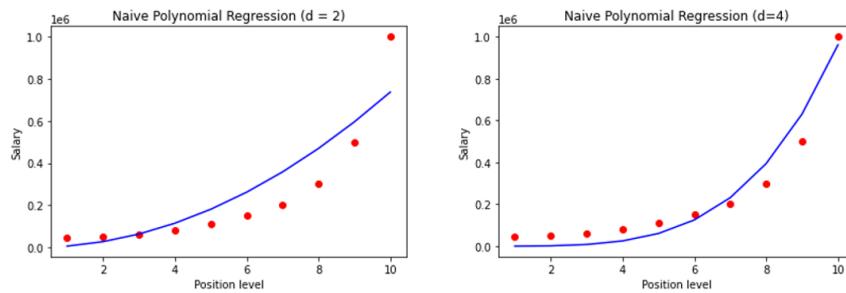
$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \varepsilon_i \text{ for } i = 1, 2, \dots, n \implies \mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon,$$

where  $\mathbf{x}, \mathbf{y}, \varepsilon, \boldsymbol{\theta} \in \mathbb{R}^{n \times 1}$ ,  $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{n \times 3}$ ,  $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^\top \in \mathbb{R}^{3 \times 1}$ .

So,  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon$  is the general form of a linear model, where

- a.  $\mathbf{y}$  is called the response such that  $E(\mathbf{y}) = \mathbf{X}\theta$ .
- b.  $\mathbf{x}$  is called the explanatory variable, the regressor, or the covariates.
- c.  $\mathbf{X}$  is called the design matrix.
- d. A least square estimate of  $\theta$ :

$$\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\hat{\theta}_0, \hat{\theta}_1)^\top = (10.73, 0.8726)^\top.$$



**Figure 6.4:** Estimated embedded linear models  $\hat{y} = 2.25 \times 10^5 - 1.40 \times 10^5 x + 2.09 \times 10^4 x^2$  and  $\hat{y} = 1.35 \times 10^5 - 1.43 \times 10^5 x + 6.82 \times 10^4 x^2 - 1.16 \times 10^4 x^3 + 7.09 \times 10^2 x^4$  that can predict the salary for an employee with given position number.

#### 6.2.4 Multiple multivariate linear model (MMLM)

Motivation and intuition of multiple multivariate linear model:

**Definition 6.4** (Kronecker product). Consider two  $2 \times 2$  matrices as:

$$A = \begin{pmatrix} 2 & 0 \\ 1 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 5 & -1 \\ -1 & 4 \end{pmatrix}$$

Then, Kronecker product is defined as:

$$A \otimes B = \begin{pmatrix} 2B & 0B \\ B & 3B \end{pmatrix} = \begin{pmatrix} 10 & -2 & 0 & 0 \\ -2 & 8 & 0 & 0 \\ 5 & -1 & 15 & -3 \\ -1 & 4 & -3 & 12 \end{pmatrix}.$$

### Computation of parameters for MMLM

**Theorem 6.13** (Simplification of MMLM in column and row pictures). Backdrop: Consider a realization  $\mathbf{z} = (z_1, \dots, z_n)$  of an I.I.D. random sample of size  $n$  i.e.  $(Z_1, \dots, Z_n) \rightarrow n$  I.I.D. copies of a joint random variable  $Z = (\mathbf{X}, \mathbf{Y})$  with unknown linear model  $\hat{Y}_i = E(Y_i|\mathbf{X}) = \theta_{i0} + \theta_{i1}X_1 + \dots + \theta_{id}X_d, i = 1, \dots, p \equiv Y_i = E(Y_i|\mathbf{X}) + \varepsilon = \theta_{i0} + \theta_{i1}X_1 + \dots + \theta_{id}X_d, i = 1, \dots, p$ . Now, running the linear model over  $n$  training examples namely in the data-set  $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$  as a column picture:

$$\begin{aligned} \mathbf{Y} = \mathbf{X}\Theta + \varepsilon &\implies \text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{X}\Theta) + \text{vec}(\varepsilon) \implies \\ &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} \mathbf{X}\theta_1 \\ \mathbf{X}\theta_2 \\ \vdots \\ \mathbf{X}\theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} = (I_p \otimes \mathbf{X})\text{vec}(\Theta) + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{X} & 0 & \dots & 0 & 0 \\ 0 & \mathbf{X} & \dots & 0 & 0 \\ \vdots & 0 & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \mathbf{X} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} \implies \\ y_j &= \mathbf{X}\theta_j + \varepsilon_j, j = 1, \dots, p. \end{aligned}$$

As a row picture:

$$\begin{aligned} \mathbf{Y} = \mathbf{X}\Theta + \varepsilon &\implies \text{vec}(\mathbf{Y}^\top) = \text{vec}((\mathbf{X}\Theta)^\top) + \text{vec}(\varepsilon^\top) \implies \\ &= \begin{pmatrix} \mathbf{Y}_1^\top \\ \mathbf{Y}_2^\top \\ \vdots \\ \mathbf{Y}_n^\top \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1\Theta)^\top \\ (\mathbf{X}_2\Theta)^\top \\ \vdots \\ (\mathbf{X}_n\Theta)^\top \end{pmatrix} + \begin{pmatrix} \varepsilon_1^\top \\ \varepsilon_2^\top \\ \vdots \\ \varepsilon_n^\top \end{pmatrix} = (\mathbf{X} \otimes I_p)\text{vec}(\Theta^\top) + \begin{pmatrix} \varepsilon_1^\top \\ \varepsilon_2^\top \\ \vdots \\ \varepsilon_n^\top \end{pmatrix} \\ &= \begin{pmatrix} x_{11}I_p & x_{12}I_p & \dots & x_{1d}I_p \\ x_{21}I_p & x_{22}I_p & \dots & x_{2d}I_p \\ \vdots & \vdots & \dots & \vdots \\ x_{n1}I_p & x_{n2}I_p & \dots & x_{nd}I_p \end{pmatrix} \begin{pmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1^\top \\ \varepsilon_2^\top \\ \vdots \\ \varepsilon_n^\top \end{pmatrix} \implies \\ Y_i &= (\mathbf{X}_i \otimes \Theta) + \varepsilon_i, i = 1, \dots, n. \end{aligned}$$

**Remark 6.1.** Computation and estimation of parameters for the multiple multivariate linear models can be carried with the help of simple multivariate linear models approach using a suitable decomposition technique either in row or column picture form.

### 6.3 Generalized linear models

First, we consider a model conditional probability of success for  $Y \in \{0, 1\}$  given  $X, P(Y = 1|X)$ , as:

$$\begin{aligned} \text{logit}(P(Y = 1|X)) &= \log \left( \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right) = \theta^\top X \implies \\ P(Y = 1|X) &= \frac{1}{1 + \exp(-\theta^\top X)} = \frac{\exp(\theta^\top X)}{1 + \exp(\theta^\top X)} = \sigma(\theta^\top X). \end{aligned}$$

So, in general a statistical modelling (i.e. a categorical modelling), we wish to estimate/predict a discrete class label (or a posteriori probability), with a linear argument and a known activation as:

$$\text{logit}(\hat{Y}) = \ln \left( \frac{\hat{Y}}{1 - \hat{Y}} \right) = \theta^\top X \implies \hat{Y} = \frac{1}{1 + \exp(-\theta^\top X)} = \sigma(\theta^\top X).$$

#### 6.3.1 Logistic regression: Bernoulli logit model

Algorithmic approach of logistic regression:

Step 1: First transform the space of class probability to the space of logit of the probability using the transformation

$$\text{logit}(p) = \log \left( \frac{p}{1 - p} \right) \in (-\infty, \infty), \quad p \in [0, 1].$$

Step 2: Now, we perform linear regression as the linear model

$$\text{logit}(\hat{Y}) = \text{logit}(p) = \theta^\top X$$

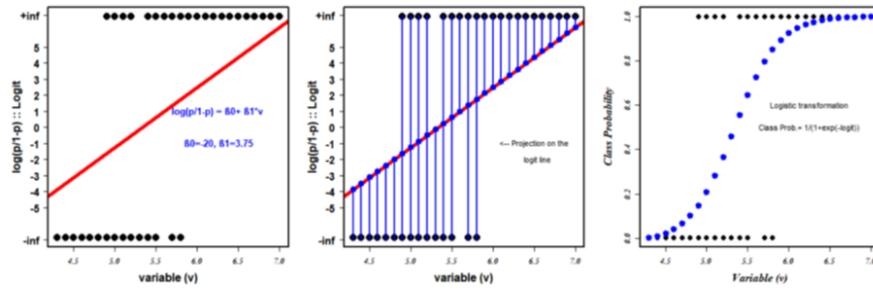
in order to maximize the likelihood of observing the data.

Step 3: Finally, we remap the space of  $\text{logit}(p)$  to class probability using the inverse of the logit as the logistic map as:

$$\hat{Y} = \hat{p} = \text{logit}^{-1}(\theta^\top X) = \frac{\exp(\theta^\top X)}{1 + \exp(\theta^\top X)} = \sigma(\theta^\top X).$$

Step 4: Further, as  $p \in [0, 1]$ , we have a decision boundary at:

$$p = 0.5 \implies \theta^\top X = \text{logit}(0.5) = \log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0.$$



**Figure 6.5:** A visualization of logistic regression via linear model of  $\text{logit}(p)$  followed by inversion map of the logit function.

### Example 6.8.

Now, running over the given dataset  $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ , we have

**Theorem 6.14.** Consider  $\mathcal{X}$  be a set of inputs.  $Y \in \{-1, 1\}$  binary outputs,  $\mathcal{S} = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$  a training set of  $n$  pairs. Estimate a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  to predict  $y$  by  $\text{sign}(f(x))$ .

*Proof.* Consider a 0/1 loss measures if a prediction is correct or not as:

$$\ell_{0/1}(y, f(x)) = I(yf(x) < 0) = \begin{cases} 0 & \text{if } y = \text{sign}(f(x)) \\ 1 & \text{otherwise.} \end{cases}$$

And, the estimate of  $f$ , we get by solving

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n \ell_{0/1}(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

The minimization problem of minimizing misclassification error with regularization is having various limitation as follows:

- a. the problem is non-smooth, and typically NP-hard to solve,

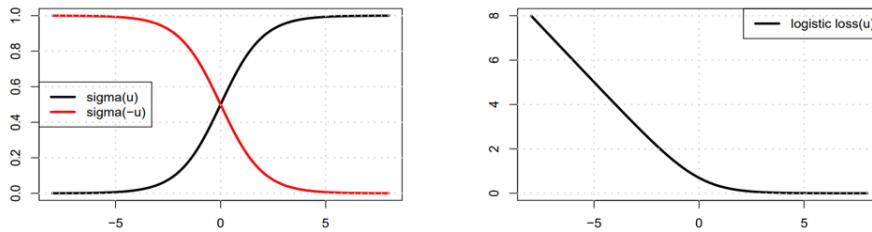
- b. the regularization has no effect since the 0/1 loss is invariant by scaling of  $f$ . In fact, no function achieves the minimum when  $\lambda > 0$ .

So, an alternative is to define a probabilistic model of  $y \in \{-1, 1\}$  parametrized by  $f(x)$ , as follows:

$$\begin{aligned} y\omega^\top x = yf(x) &= \text{logit}(E(Y = y|X = x)) = \text{logit}(P(Y = y|X = x)) \implies \\ p(y|x) &= \text{logit}^{-1}(yf(x)) = \sigma(yf(x)) = \frac{1}{1 + \exp(-yf(x))} \implies \\ \ell_{\text{logistic}}(y, f(x)) &= -\ln(p(y|x)) = \ln(1 + \exp(-yf(x))). \end{aligned}$$

A few properties of sigmoid-logistic function and logistic loss can be characterized as:

$$\begin{aligned} \sigma(u) &= \frac{1}{1 + \exp(-u)}, \sigma(-u) = 1 - \sigma(u), \sigma'(u) = \sigma(u)\sigma(-u), \\ \ell_{\text{logistic}}(u) &= \ln(1 + \exp(-u)), \ell'_{\text{logistic}}(u) = -\sigma(-u), \ell''_{\text{logistic}}(u) = \sigma'(u). \end{aligned}$$



**Figure 6.6:** Plots of sigmoid-logistic function and logistic loss.

Now, the modified estimation problem can be posed as:

$$\begin{aligned} \hat{f} &= \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n \ell_{\text{logistic}}(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right) \\ &= \arg \min_{f \in \mathcal{H}} \left( \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i f(x_i))) + \lambda \|f\|_{\mathcal{H}}^2 \right). \end{aligned}$$

By the representation theorem, the solution of the unconstrained problem can be expanded as:

$$\hat{f}(x) = \sum_{i=1}^n \alpha_i k(x_i, x) \implies (\hat{f}(x_1), \dots, \hat{f}(x_n))^\top = K\alpha, \|f\|_{\mathcal{H}} = \alpha^\top K\alpha.$$

To estimate  $\alpha$ , we need to solve the following:

$$\begin{aligned}\hat{\alpha} &= \arg \min_{\alpha \in \mathbb{R}} \left( \frac{1}{n} \sum_{i=1}^n \ln (1 + \exp(-y_i(K\alpha)_i)) + \lambda \alpha^\top K\alpha \right) \\ &\approx \arg \min_{\alpha \in \mathbb{R}} \left( \frac{1}{n} (K\alpha - z)^\top W (K\alpha - z) + \lambda \alpha^\top K\alpha \right).\end{aligned}$$

The above convex quadratic approximation (with respect to an affine transformation  $z = K\alpha_0 - W^{-1}Py$ ) is as follows:

$$\begin{aligned}\nabla J(\alpha) &= \frac{1}{n} KP(\alpha)y + \lambda K\alpha, P(\alpha) = \text{diag}(P_1(\alpha), \dots, P_n(\alpha)), \\ P_i(\alpha) &= \ell'_{logistic}(y_i f(x_i)) = -\sigma(-y_i f(x_i)) \\ \nabla^2 J(\alpha) &= \frac{1}{n} KW(\alpha)K + \lambda K, W(\alpha) = \text{diag}(W_1(\alpha), \dots, W_n(\alpha)), \\ W_i(\alpha) &= \ell''_{logistic}(y_i f(x_i)) = \sigma(y_i f(x_i))\sigma(-y_i f(x_i)).\end{aligned}$$

So, the estimate of  $\alpha$  is computed as:

$$\begin{aligned}\nabla J_{quadratic}(\alpha) &= \frac{2}{n} KW(K\alpha - z) + 2\lambda K\alpha = \frac{2}{n} K ((WK + \lambda nI) - Wz) = 0 \\ \implies \alpha - (WK + \lambda nI)^{-1} Wz &\in \text{Ker}(K) \implies \alpha = (WK + \lambda nI)^{-1} Wz + \varepsilon, \\ \varepsilon \in \text{Ker}(K) &\implies \hat{\alpha} = (WK + \lambda nI)^{-1} Wz \text{ with } K\varepsilon = 0.\end{aligned}$$

□

### 6.3.2 SoftMax regression: a multinomial logit model

## 6.4 Future note: control, discovery models, and generation

In order to capture discrete pattern in the response variable, we need to introduce some kind of non-linear function on it (e.g. logit, multinomial logit etc.), that leads to generalized linear models (i.e. a model having known & non-linear activation function with linear argument e.g.

logistic regression, multinomial logistic regression etc.). Furthermore, discovery of patterns within input space with a few or no labels (i.e. prediction of labels or response variable via representation learning) lead to discovery models. At the end, the linear models, the generalized linear models and the discovery models can be unified by universal approximation of neural networks.

### **A machine learning system**

Step 1: Methods:-mathematical modelling with scientific meaning, estimation/optimization, optimization algorithm/ numerical methods, generalization/concentration inequalities, inference/prediction, and numerical methods etc.

Step 2: Techniques:-machine learning algorithms for decision making, control, discovery and generation (e.g. regression, classification, neural networks etc.).

Step 3: Tools: Implementation in computing devices using some platform e.g. Scikit Learn (for machine learning algorithms other than neural networks), TenFlow, PyTorch (neural networks), Keras (deep learning).

Step 4: Deployment in the real world system e.g. a self deriving car etc.

#### **6.4.1 A non-linear inverse problem**

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.

## **Acknowledgements**

---

I dedicate this text to my wife, my son, my parents, my grand-parents, my in-law-parents, my teachers and all those who help and inspire me, because through you I become whole.

This text developed out from lectures given at Indian Institute of Information Technology Dharwad in 2022 and 2023 for engineering students specializing in computation mathematics, probabilistic modelling, statistical estimation & inference.

I thank to my colleagues at IIIT Dharwad: Somen, Arun and Jagadish for their help, support and valuable discussions.

Last but not the least; I would like to convey my special thanks to my dear wife, my dear son, grandfather, parents, teachers, brothers and sisters who tolerated the onslaughts of time to bring me up to this position and for their blessings, continuous encouragement and moral support.

## References

---

- D. Bertsekas and J.N. Tsitsiklis. Introduction to probability. *Athena Scientific*, 2008.
- M. S. Kay. Intuitive probability and random processes using MATLAB®. *Springer Science & Business Media*, 2006.
- R. J. Freund and P. D. Minton. Regression methods : a tool for data analysis. *M. Dekker, New York*, 1979.
- P. E. Pfeiffer and D. A. Schum. Introduction to APPLIED PROBABILITY. *ACADEMIC PRESS New York and London*, 1973.
- W. Feller. An introduction to probability theory and its applications. *Wiley. ISBN: 0471257087*, 1968.
- A. Tsun. Probability & Statistics with Applications to Computing. *online notes*, 2020.
- R. V. Handel. Probability and Random Processes. *ORF 309/MAT 380 Lecture Notes, Princeton University*, 2006.