# INTRODUCTION TO STATISTICS

## THEORY, METHODS, AND APPLICATIONS

AUTHORED BY

JONAS PETERSEN

December 4, 2024

This page is intentionally left blank

# Contents

# CHAPTER 1

## Preface

Statistics is a mathematical discipline that use probability theory (which in turn require set theory) to extract insights from information (data). Probability theory is a branch of pure mathematics – probabilistic questions can be posed and solved using axiomatic reasoning, and therefore there is one correct answer to any probability question. Statistical questions can be converted to probability questions by the use of probability models. Given certain assumptions about the mechanism generating the data, statistical questions can be answered using probability theory. This highlights the dual nature of statistics, comprised of two integral parts.

1. The first part involves the formulation and evaluation of probabilistic models, a process situated within the realm of the philosophy of science. This phase grapples with the foundational aspects of constructing models that accurately represent the problem at hand.

2. The second part concerns itself with extracting answers after assuming a specific model. Here, statistics becomes a practical application of probability theory, involving not only theoretical considerations but also numerical analysis in real-world scenarios.

This duality underscores the interdisciplinary nature of statistics, bridging the gap between the conceptual and the applied aspects of probability theory. Although probabilities are well defined (see chapter 3), their interpretation is not defined beyond their definition. This ambiguity has given birth to two competing interpretations of probability, leading to two competing branches of statistics; Frequentist and Bayesian Statistics. This book aims to explain how these competing branches of statistics fit together as well as providing a non-exhaustive presentation of some of the methods within both branches. The philosophy of the book is rather straight to the point, but with a lot of examples both big and small. Some of these are anonymized versions of projects from industry. The books is split into three parts; introduction (part i), Frequentist statistics (part ii) and Bayesian statistics (part iii).

## 1.1 ACKNOWLEDGEMENTS

The philosophy of the book is similar to [1], a few exercises from [2] used as examples, the idea of phrasing decision theory as "Robot vs Nature" from [3] and the review of probability theory is inspired by [4].

# Part I

# INTRODUCTION

# CHAPTER 2

## Introduction to Set Theory

Set theory is a fundamental branch of mathematical logic that provides a foundation for much of mathematics, including probability theory. At its core, set theory deals with the concept of a set, which is a collection of distinct objects or elements. In this introduction, the essential properties and operations of sets are explored in order to lay the groundwork for the axiomatic formation of probability theory and statistics.

**Definition 1** (Membership). *In set theory, the membership relation between an object $o$ and a set $A$ is fundamental. $o \in A$ denotes that $o$ is an element or member of $A$.*

**Definition 2** (Set). *A set is a collection of distinct objects, considered as an object in its own right. Sets are typically denoted using curly braces $\{\}$ and can be described in two primary ways:*

1. *By listing its elements separated by commas, e.g., $A = \{a_1, a_2, a_3\}$.*
2. *By specifying a characterizing property of its elements, e.g., $A = \{x \mid x \text{ is a natural number}\}$.*

*Sets can also be illustrated graphically, as shown in Figure 1.*

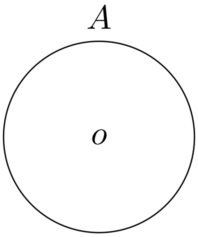

Figure 1: The graphical representation of a generic set $A$ with generic elements $o$.

**Definition 3** (Subset). *A set $A$ is called a subset of a set $B$, denoted $A \subseteq B$, if every element of $A$ is also an element of $B$. Formally, $A \subseteq B$ if $\forall x \in A, x \in B$. By this definition, a set is always a subset of itself.*

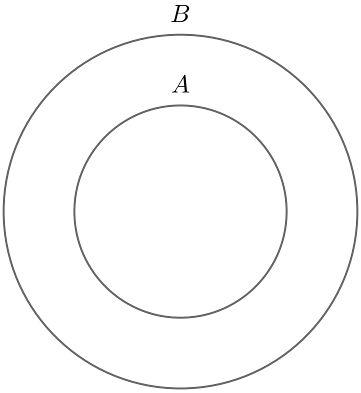

Figure 2: The graphical representation of $A \subseteq B$.

**Definition 4** (Proper Subset). *A set $A$ is called a proper subset of a set $B$, denoted $A \subset B$, if $A \subseteq B$ and $A \neq B$. This means that $A$ is a subset of $B$ but $A$ is not equal to $B$; there is at least one element in $B$ that is not in $A$.*

**Example 2.1.**

*Suppose $A =$ {🍌, 🍎, 🍆}, then {🍌, 🍎} and {🍎} are proper subsets* *of $A$, meaning* {🍌,🍎},{🍎} $\subset A$. {🍌,🥕}*, on the other hand, is not a subset of $A$, meaning* {🍌,🥕} $\not\subset A$.

---

**Example 2.2.**

🍌, 🍎, *and* 🍆 *are members (elements) of the set* {🍌,🍎,🍆}*, but are not subsets of it; and in turn, the subsets, such as* {🍌}*, are not members of the set* {🍌,🍎,🍆}.

---

**Definition 5** (Empty Set). *The empty set, denoted by $\emptyset$ or $\{\}$, is the set that contains no elements.*

**Definition 6** (Universal Set). *The universal set, denoted by $\Omega$, is the set that contains all the objects or elements under consideration in a particular discussion or problem. It is the largest set in the context of a given study.*

**Definition 7** (Closure). *A set $A$ is said to be closed under a certain operation if, for every pair of elements $x$ and $y$ in $A$, the result of applying the operation to $x$ and $y$ is also in $A$.*

**Definition 8** (Union). *The union of sets $A$ and $B$, denoted by $A \cup B$, is defined as the set containing all elements that are in $A$ or $B$ (or both). Figure 3 provide a graphical representation of $A \cup B$.*

**Definition 9** (Intersection). *The intersection of sets $A$ and $B$, denoted by $A \cap B$, is defined as the set containing all elements that are common to both $A$ and $B$. Figure 4 provide a graphical representation of $A \cap B$.*

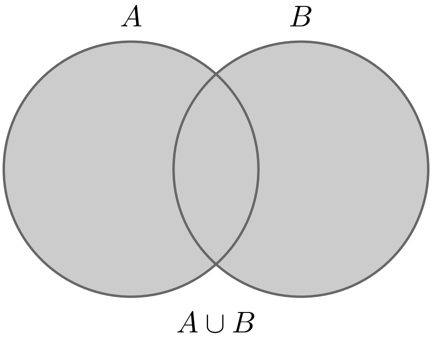

Figure 3: The figure show the union of sets $A$ and $B$. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

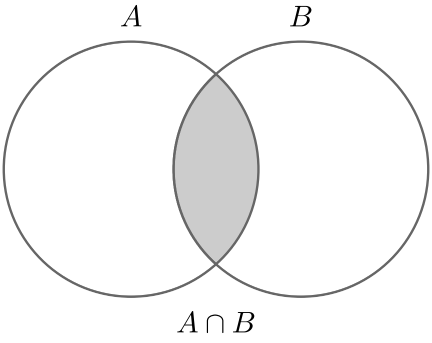

Figure 4: The figure show the intersection of sets $A$ and $B$. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

**Definition 10** (Disjoint). *Two sets $A$ and $B$ are said to be disjoint if their intersection is the empty set, i.e., $A \cap B = \emptyset$. Figure 5 provide a graphical representation of $A \cap B = \emptyset$.*

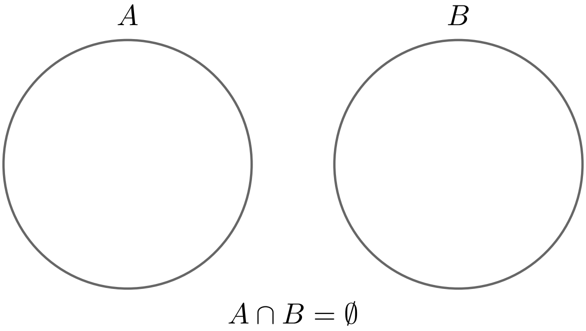

Figure 5: The figure show the case where the intersection of sets $A$ and $B$ is the empty set. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

**Definition 11** (Complementation). *The complement of set $A$, denoted by $A^c$, is defined as the set containing all elements in the universal set $\Omega$ that are not in $A$. Figure 6 provide a graphical representation of $(A \cap B)^c$.*

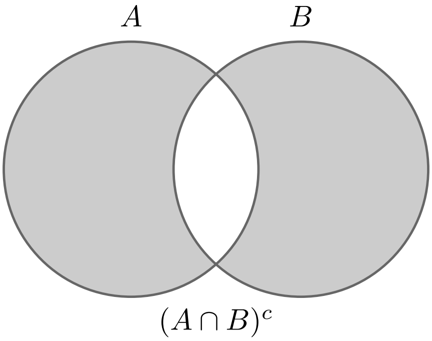

Figure 6: The figure show the complementary of the intersection of sets $A$ and $B$. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

**Definition 12** (Difference). *The difference between set $A$ and $B$, denoted by $A \setminus B = A \cap B^c$, is defined as the set containing all elements in $A$ that are not in $B$. Figure 7 provide a graphical representation of $A \setminus B$ and $B \setminus A$.*

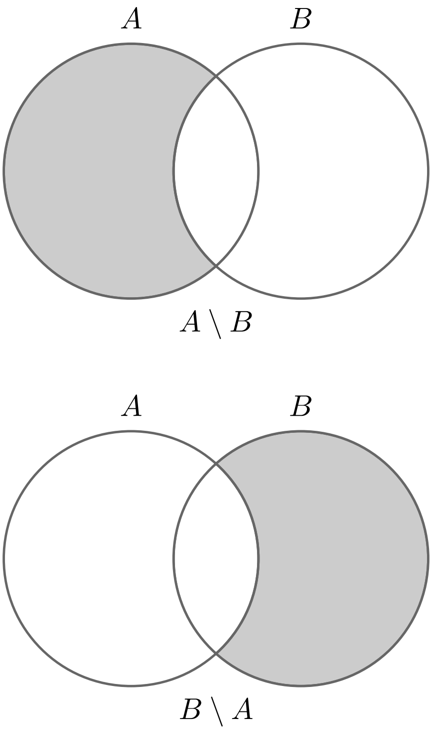

Figure 7: (left) show $A$ minus $B$ and (right) show $B$ minus $A$. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

**Definition 13** (Power Set). *The power set of a set $A$, denoted by $2^A$, is defined as the set containing all possible subsets of $A$, including $A$ itself and the empty set.*

**Example 2.3.**

*Suppose $A = \{a_1, a_2, a_3\}$, then*

$$\begin{aligned} 2^A = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \\ \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}\}. \end{aligned} \tag{1}$$

**Definition 14** (Symmetric Difference). *The symmetric difference of sets $A$ and $B$, denoted by $A \Delta B$, is defined as the set containing all elements that are in either $A$ or $B$ but not in both, meaning $A \Delta B = (A \cap B)^c$. Figure 6 show the symmetric difference between sets $A$ and $B$.*

**Definition 15** (Finite and Infinite Unions). *For a collection $\{A_i\}$, the union is denoted by $\bigcup_i A_i$ and is defined as the set containing all elements that are in at least one of the sets $A_i$.*

**Definition 16** (Partition). *A collection of non-empty subsets $\{A_i\}$ of a set $A$ is called a partition of $A$ if the following conditions are satisfied:*

1. *The subsets $A$ are pairwise disjoint, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$.*
2. *The union of all subsets $A_i$ is equal to the set $A$, i.e., $\bigcup_{i \in I} A_i = A$.*

*Figure 8 provide a graphical representation of the set $A = \{A_1, A_2, A_3\}$ where $A_j$ are partitions.*

**Definition 17** (Finite and Infinite Intersections). *For a collection $\{A_i\}$, the intersection is denoted by $\bigcap_i A_i$ and is defined as the set containing all elements that are common to all sets $A_i$.*

**Definition 18** (Cartesian Product). *The Cartesian product of sets $A$ and $B$, denoted by $A \times B$, is defined as the set containing all ordered pairs $(a, b)$, where $a$ is in $A$ and $b$ is in $B$.*

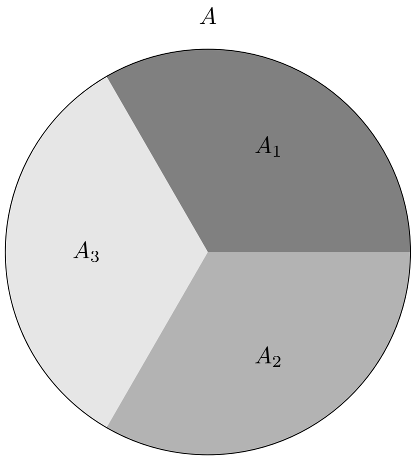

Figure 8: The figure show $A = \{A_1, A_2, A_3\}$ where $A_j$ are partitions.

**Example 2.4.**

*Suppose* $A = \{a_1, a_2\}$ *and* $B = \{b_1, b_2, b_3\}$*, then*

$$\begin{aligned} A \times B = \{&(a_1, b_1), (a_1, b_2), (a_1, b_3), \\ &(a_2, b_1), (a_2, b_2), (a_2, b_3)\} \end{aligned} \tag{2}$$

# CHAPTER 3

# Introduction to Probability Theory

Probability theory aims to provide a mathematical framework for analyzing random experiments, where outcomes cannot be predicted with certainty beforehand. Its objective is to systematically study and understand the potential outcomes of these experiments.

**Definition 19** (Sample Space)**.** *The sample space, denoted by $\Omega$, represents the set of all possible outcomes of a random experiment. It encompasses every conceivable result that could occur, serving as the foundation for analyzing probabilities associated with different outcomes.*

**Definition 20** (Event)**.** *An event, $E$, is a subset of the sample space, denoted by $E \subseteq \Omega$, that corresponds to a specific collection of possible outcomes in a random experiment. Events may consist of single or multiple outcomes and are defined by the occurrence or non-occurrence of particular conditions.*

**Example 3.1.**

*Consider the roll of a fair six-sided die. The sample space for this experiment is given by $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. $E = \{$⚁, ⚃, ⚅$\}$, is the event of rolling an even number.*

**Definition 21** (Event Space)**.** *The set containing all valid possible events for a random experiment is referred to as the event space, $\mathcal{F}$. The notion of "all valid possible events for a random experiment" is formally defined by requiring $\mathcal{F}$ to be a $\sigma$-algebra satisfying the following properties:*

1. *$\mathcal{F}$ is the set of all subsets of the sample space $\Omega$, including the empty set $\emptyset$ and $\Omega$ itself, along with various combinations of outcomes.*

2. *Closure under complementation: If $E$ is in the $\sigma$-algebra ($E \in \mathcal{F}$), then its complement $E^c$ is also in the $\sigma$-algebra.*

3. *Closure under countable union and intersection: If the events $E_1, E_2, E_3, \ldots$ are in the $\sigma$-algebra ($E_i \in \mathcal{F}$ for all $i$), then their countable union $\bigcup_{i=1}^{\infty} E_i$ and intersection $\bigcap_{i=1}^{\infty} E_i$ are also in the $\sigma$-algebra.*

*In the case where the outcomes of the random experiment can take discrete values, these properties are sufficient. However, in the case where the outcomes are continuous, $\mathcal{F}$ is required to be a Borel $\sigma$-algebra, meaning it must further fulfill the closure property under countable intersection with open sets. This ensures that $\mathcal{F}$ contains all sets that can be formed by taking unions, intersections, and complements of open sets, which are essential for defining probabilities in continuous spaces.*

**Example 3.2.**

*For the roll with the fair die considered in example 3.1, the sample space is $\Omega = \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\}$ and the event space (the set of all possible events) is given by*

$$\begin{aligned}\mathcal{F} &= \{\emptyset, \{\text{⚀}\}, \{\text{⚀}, \text{⚂}\}, \{\text{⚂}\}, \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}\}, \{\text{⚄}\}, \ldots\} \\ &= 2^{\Omega}.\end{aligned} \tag{3}$$

**Definition 22** (Measurable Space)**.** *The pair $(\Omega, \mathcal{F})$ is called a measurable space.*

Probability can loosely be defined [4] as a measure of the size of an event (a set) relative to the sample space (another set), meaning it is a function that operates on an event (a set). In particular the probability measure maps any valid event, i.e. any $E \in \mathcal{F}$, to a number between 0 and 1, representing the relative size of the event to the sample space.

**Definition 23** (Probability Measure). *A Probability measure, $\mathbb{P}$, is a set function defined on a measurable space (definition 22) $(\Omega, \mathcal{F})$*

$$\mathbb{P} : \mathcal{F} \mapsto [0,1] \tag{4}$$

*that obey [5] axioms 1-3.*

**Axiom 1** (Non-negativity). *For any event $E \in \mathcal{F}$, the probability measure $\mathbb{P}(E)$ is non-negative, satisfying*

$$\mathbb{P}(E) \geq 0 \quad \forall E \in \mathcal{F}. \tag{5}$$

**Axiom 2** (Normalization). *The probability of the universal set $\Omega$ is 1, satisfying*

$$\mathbb{P}(\Omega) = 1. \tag{6}$$

**Axiom 3** (Additivity). *For any countable sequence of mutually exclusive events $E_1, E_2, \ldots \in \mathcal{F}$, the probability of their union is the sum of their individual probabilities, such that*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) \quad \forall E_i \in \mathcal{F} \textit{ where } \bigcap_{i=1}^{\infty} E_i = \emptyset. \tag{7}$$

Together, the probability measure, the sample space and the algebra form the tuple $(\Omega, \mathcal{F}, \mathbb{P})$ which define what a probability space. The non-negativity and normalization axioms are largely matters of convention, although it is nontrivial that probability measures take at least the two values 0 and 1, and that they have a maximal value (unlike various other measures, such as length, volume, and so on, which are unbounded). The axioms are supplemented by two definitions.

**Definition 24** (Conditional Probability). *For events $E_1$ and $E_2$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(E_2) > 0$, the conditional probability of $E_1$ given $E_2$ is defined viz*

$$\mathbb{P}(E_1|E_2) \equiv \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2)}, \tag{8}$$

*where $\mathbb{P}(E_1, E_2) = \mathbb{P}(E_1 \cap E_2)$ to ease the notation.*

**Definition 25** (Independence). *Events $E_1$ and $E_2$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be conditionally independent if*

$$\mathbb{P}(E_1, E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2). \tag{9}$$

From axioms 1-3 and definitions 8 and 9, the chain rule, the concept of marginalization, conditional independence and the law of total probability can be derived.

**Theorem 1** (Chain Rule). *Given $\{E_1, E_2, \ldots, E_n\} \subseteq \mathcal{F}$ denotes a set of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the chain rule for this set of events can be written*

$$\mathbb{P}(E_1, \ldots E_n) = \mathbb{P}(E_1) \prod_{j=2}^{n} \mathbb{P}(E_j | E_1, \ldots E_{j-1}). \tag{10}$$

*Proof.* From the definition of conditional probability in equation (8)

$$\mathbb{P}(E_1, E_2, \ldots, E_n) = \mathbb{P}(E_1 | E_2, \ldots, E_n)\mathbb{P}(E_2, \ldots, E_n). \tag{11}$$

Using the definition of conditional probability again

$$\mathbb{P}(E_2, \ldots, E_n) = \mathbb{P}(E_2 | \ldots, E_n)\mathbb{P}(\ldots, E_n). \tag{12}$$

Continuing in this way, equation (10) follows. □

Equation (10) illustrates how to decompose the joint probability of multiple events into a product of conditional probabilities. The idea is to calculate the probability of each event in the sequence conditioned on the occurrence of the previous events in the chain. The chain rule is particularly powerful when dealing with complex systems where events may be interdependent. It allows breaking down joint probabilities into more manageable conditional probabilities, making it easier to analyze and model intricate relationships between events. Whether in the context of statistical modeling or machine learning, the chain rule plays a key role in calculating the joint probability of multiple events and provides a foundation for more advanced probabilistic reasoning.

**Theorem 2** (Bayes theorem). *For events $E_1, E_2, E_3 \in \mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Bayes theorem can be formulated viz*

$$\mathbb{P}(E_1|E_2, E_3) = \frac{\mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2, E_3)}. \tag{13}$$

*Proof.* Bayes theorem follows directly from applying the chain rule and applying the concept of symmetry viz

$$\begin{aligned} \mathbb{P}(E_1, E_2, E_3) &= \mathbb{P}(E_1|E_2, E_3)\mathbb{P}(E_2, E_3) \\ &= \mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_3) \end{aligned} \tag{14}$$

from which

$$\mathbb{P}(E_1|E_2, E_3) = \frac{\mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2, E_3)} \tag{15}$$

which is Bayes theorem. □

**Theorem 3** (Law of Total Probability). *Let $\{E_1, E_2, \dots E_n\}$ be a partition of the sample space $\Omega$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then for any $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, E_i). \tag{16}$$

*In continuous cases, the summation is replaced by integration.*

*Proof.* Consider an event $A \subseteq \Omega$ and a partition $\{E_1, E_2, \dots E_n\}$ of $\Omega$ such that $\cup_i E_i = \Omega$. For mutually exclusive events (which a partition by definition is), finite additivity can be used such that

$$\sum_i \mathbb{P}(A, E_i) = \mathbb{P}(\bigcup_i (A, E_i)). \tag{17}$$

$\bigcup_i (A, E_i)$ is the union of all intersections between $A$ and the $E$'s. However, since the $E$'s form a partition of $\Omega$, they together form $\Omega$ and the intersection between $\Omega$ and $A$ is $A$, meaning

$$\begin{aligned}\bigcup_i (A, E_i) &= (A, \bigcup_i E_i)\\ &= (A, \Omega)\\ &= A.\end{aligned} \tag{18}$$

Combining equations (17)-(18) then yields

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, E_i). \tag{19}$$

□

**Example 3.3.**

*Consider the roll of a fair six-sided die. The sample space for this experiment is given by* $\Omega =$ {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. *Let* $E_1 =$ {⚁, ⚃, ⚅} *and* $E_2 =$ {⚃} *be two events, then from equation* (8)

$$\begin{aligned}\mathbb{P}(E_1|E_2) &= \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2)}\\ &= 1\end{aligned} \tag{20}$$

*where* $\mathbb{P}(E_1, E_2) = \frac{1}{6}$ *since* $E_1, E_2 = E_1 \cap E_2 = E_2 =$ {⚃} *is one of* 6 *possible values and* $\mathbb{P}(E_2) = \frac{1}{6}$. *Intuitively this makes sense because* $E_2$ *is a set with one member and since* $E_2$ *is known, the outcome of the experiment is known with certainty in this case.*

**Definition 26** (Random Variable). *A random variable* $X$ *is a function*

$$X : \Omega \mapsto \Omega_X \tag{21}$$

*that maps outcomes from a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$, *where* $\Omega_X$ *is the codomain of* $X$ *and* $\mathcal{F}_X$ *is a* $\sigma$*-algebra on* $\Omega_X$. *The* $\sigma$*-algebra* $\mathcal{F}_X$ *ensures that* $X$ *is measurable, meaning that for any set* $x \in \mathcal{F}_X$, *the preimage* $X^{-1}(x)$ *must belong to* $\mathcal{F}$. *Formally, this can be written as*

$$X^{-1}(x) = \{\omega \in \Omega | X(\omega) = x\} \in \mathcal{F} \quad \forall x \in \mathcal{F}_X. \tag{22}$$

*Random variables are classified as either discrete or continuous, based on the discrete or continuous nature of their sample space. Discrete random variables have countable sample spaces, while continuous random variables have uncountable sample spaces, often modeled as intervals on the real line. The role of random variables is to provide a numerical representation of the outcomes of a random experiment, allowing quantification and analysis of the likelihood of different numerical outcomes.*

**Definition 27** (Expected value). *Let $X$ be a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the expected valued of $X$, denoted by $\mathbb{E}[X]$, is defined by the Lebesgue integral [6]*

$$\mathbb{E}_X[X] \equiv \int_\Omega X(\omega) d\mathbb{P}(\{\omega\}). \tag{23}$$

**Theorem 4** (Non-negativity of expected value). *If $X \geq 0$ for a random variable $X$, then $\mathbb{E}_X[X] \geq 0$.*

**Theorem 5** (Linearity of expected value). *The expectation is a linear operator meaning $\mathbb{E}_X[a + X] = a + \mathbb{E}_X[X]$ and $\mathbb{E}_X[aX] = a\mathbb{E}_X[X]$ for any constant $a$.*

**Theorem 6** (The law of the unconscious statistician). *The law of the unconscious statistician generalize the expectation of a random variable to the expectation of a function $g : \Omega \mapsto \mathbb{R}$ of a random variable $X(\omega) \in \Omega_X \quad \forall \omega \in \Omega$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

$$\mathbb{E}[g(X)] \equiv \int_\Omega g(X(\omega)) d\mathbb{P}(\{\omega\}). \tag{24}$$

**Definition 28** (Variance). *Let $X$ be a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the variance of $X$, denoted by $Var[X]$, is defined viz*

$$\begin{aligned} Var[X] &\equiv \mathbb{E}_X[(X - \mathbb{E}_X[X])^2] \\ &= \mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2. \end{aligned} \tag{25}$$

**Theorem 7** (Non-linearity of variance). *The variance is a non-linear opeator, where $Var[a + X] = Var[X]$ and $Var[aX] = a^2 Var[X]$ for any constant $a$.*

**Definition 29** (Image Measure). *Let $X : \Omega \mapsto \Omega_X$ be a random variable that maps from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Omega_X, \mathcal{F}_X)$. Then [7]*

$$\mathbb{P} \circ X^{-1} : \mathcal{F}_X \mapsto [0, 1] \tag{26}$$

*defines a probability measure on $(\Omega_X, \mathcal{F}_X)$. $\mathbb{P} \circ X^{-1}$ is called the image measure or the push forward measure of $\mathbb{P}$.*

**Definition 30** (Probability Mass Function). *In case of a discrete random variable $X : \Omega \mapsto \Omega_X$ that maps from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Omega_X, \mathcal{F}_X)$, the image measure is defined as the probability mass function*

$$\begin{aligned} p(X = x) &\equiv \mathbb{P} \circ X^{-1}(x) \\ &= \mathbb{P}(X^{-1}(x)). \end{aligned} \tag{27}$$

*According to axioms 1-3 $\sum_{all\ x} p(X = x) = 1$ and $p(X = x) \geq 0 \quad \forall x \in \Omega_X$.*

**Theorem 8** (Expected value of discrete random variable). *The expected value of a discrete random variable $X$ with probability mass function $p$ can be written*

$$\mathbb{E}_X[X] = \sum_i x_i p(X = x_i). \tag{28}$$

**Definition 31** (Probability Density Function). *Let $X : \Omega \mapsto \Omega_X$ be a continuous random variable that maps from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Omega_X, \mathcal{F}_X)$. Probabilities are described using a probability density function (PDF)*

$$f : \Omega_X \mapsto \mathbb{R}_{\geq 0}, \tag{29}$$

*which is related to the probability measure viz*

$$\mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\}) = \int_{-\infty}^{x} f(X = t)\, dt, \tag{30}$$

*where* $\int_{-\infty}^{\infty} f(X=t)dt = 1$ *and* $f(X=x)=0$ *for any individual point* $x \in \mathbb{R}$.

**Theorem 9** (Expected value of continuous random variable). *The expected value of a continuous random variable* $X$ *with probability density function* $f$ *can be written*

$$\mathbb{E}_X[X] = \int_{\Omega_X} x f(X=x)dx. \tag{31}$$

**Theorem 10** (Total expectation). *The expectation of a random variable* $X$ *can be expressed in terms of another random variable* $Y$ *viz*

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y=y]], \tag{32}$$

*where the subscript specify the probability distribution the expectation is with respect to.*

*Proof.*

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} dx x f(X=x) \\
&= \int_{\Omega_Y} dy \int_{\Omega_X} dx x f(X=x, Y=y) \\
&= \int_{\Omega_Y} dy f(Y=y) \int_{\Omega_X} dx x f(X=x|Y=y) \\
&= \int_{\Omega_Y} dy f(Y=y) \underbrace{\int_{\Omega_X} dx x f(X=x|Y=y)}_{=\mathbb{E}_{X|Y}[X|Y=y]} \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y=y]].
\end{aligned} \tag{33}$$

□

**Theorem 11** (Expectation of product of independent random variables). *Let* $X(\omega) \in \Omega_X$ *and* $Y(\omega) \in \Omega_Y$ *be independent continuous random variables defined on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, *such that* $f(X=x, Y=y) = f(X=x)f(Y=y)$, *then* $\mathbb{E}_{XY}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]$.

*Proof.*

$$\begin{aligned}\mathbb{E}_{XY}[XY] &= \int_{\Omega_X}\int_{\Omega_Y} xyf(X=x,Y=y)dxdy \\ &= \int_{\Omega_X} xf(X=x)dx \int_{\Omega_Y} yf(Y=y)dy \\ &= \mathbb{E}_X[X]\mathbb{E}_Y[Y]\end{aligned} \tag{34}$$

□

**Definition 32** (Covariance). *Let $X$ and $Y$ be a real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the covariance of $X$ and $Y$, denoted by $Cov[X,Y]$, is defined viz*

$$\begin{aligned}Cov[X,Y] &= \mathbb{E}_{XY}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y],\end{aligned} \tag{35}$$

**Theorem 12** (Covariance of independent random variables). *For independent random variables $X \in \Omega_X$ and $Y \in \Omega_Y$ the covariance is given by $Cov[X,Y] = 0$.*

*Proof.* Using $\mathbb{E}_{XY}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]$ (theorem 11) in definition 32 yield $\mathrm{Cov}[X,Y] = 0$. □

**Definition 33** (Correlation). *Let $X$ and $Y$ be real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The correlation between $X$ and $Y$, denoted by $Corr[X,Y]$, is defined as*

$$\begin{aligned}Corr[X,Y] &= \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \\ &= \frac{\mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{(\mathbb{E}_X[X^2]-\mathbb{E}_X[X]^2)\,(\mathbb{E}_Y[Y^2]-\mathbb{E}_Y[Y]^2)}}.\end{aligned} \tag{36}$$

Correlation and covariance are both measures of the relationship between two random variables. While covariance indicates the extent to which two variables change together, correlation provides a standardized measure of this relationship, taking into account the scales of the variables. In particular, the correlation between two variables, denoted by $\mathrm{Corr}[X, Y]$, is the covariance of $X$ and $Y$ divided by the product of their standard deviations. This normalization makes correlation a unitless quantity that ranges between -1 and 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship. In essence, correlation provides a more interpretable measure of the strength and direction of the linear association between two variables compared to covariance.

**Definition 34** (Change of Variables for PDFs). *Let $X$ be a continuous random variable with probability density function (PDF) $f(X = x)$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose $Y = g(X)$ is a continuous and differentiable function of $X$, and let $g^{-1}$ denote the inverse function of $g$. If $Y = g(X)$ and the inverse function $g^{-1}$ exists and is differentiable, the PDF of the random variable $Y$, denoted $f(Y = y)$, can be obtained by the change of variables formula viz [1]*

$$f(Y = y) = f(X = g^{-1}(y)) \left| \frac{d}{dY} \left( g^{-1}(Y) \right) \right|_{Y=y}. \quad (37)$$

**Example 3.4.**

*Let $X$ be a continuous random variable with PDF $f(X = x)$, and let $Y = g(X) = aX + b$, where $a \neq 0$ and $b$ are constants. The inverse function is given by*

$$g^{-1}(y) = \frac{y - b}{a} \quad (38)$$

*Using definition 34*

$$\begin{aligned} f(Y=y) &= f\left(X = g^{-1}(y)\right) \left| \frac{d}{dY}\left(g^{-1}(Y)\right) \right|_{Y=y} \\ &= f\left(X = \frac{y-b}{a}\right) \left| \frac{d}{dY}\left(\frac{Y-b}{a}\right) \right|_{Y=y} \\ &= f\left(X = \frac{y-b}{a}\right) \left| \frac{1}{a} \right| . \end{aligned} \tag{39}$$

*Thus, the PDF of Y is*

$$f(Y=y) = \frac{1}{|a|} f\left(X = \frac{y-b}{a}\right). \tag{40}$$

---

**Example 3.5.**

Let $X = \ln\left(\frac{Y}{1-Y}\right)$ be a continuous random variable with PDF $f(X=x) \propto$const. The inverse function is given by

$$g^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{41}$$

*Using definition 34*

$$\begin{aligned} f(Y=y) &= f\left(X = g^{-1}(y)\right) \left| \frac{d}{dY}\left(g^{-1}(Y)\right) \right|_{Y=y} \\ &= const \cdot \frac{1-Y}{Y}\left(\frac{1}{1-Y} + \frac{Y}{(1-Y)^2}\right)\Bigg|_{Y=y} \\ &= const \cdot Y^{-1}(1-Y)^{-1}|_{Y=y} \\ &= Beta(Y=y|a=0, b=0). \end{aligned} \tag{42}$$

---

**Definition 35** (Error-Propagation). *Let $X_1 \ldots, X_n$ be continuous random variables with means $\mathbb{E}[X_1] \ldots, \mathbb{E}[X_n]$ and variances denoted $Var[X_1] \ldots, Var[X_n]$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider a function $g(X_1, \ldots, X_n)$ of these random* *variables, $g(X_1,\ldots,X_n)$. The variance of g, which quantifies the uncertainty in g due to the uncertainties in the $X_i$, can be written*

$$\begin{aligned} Var[g(X_1\ldots,X_n)] &\equiv \mathbb{E}\left[(g(X_1\ldots,X_n)-\mathbb{E}[g(X_1\ldots,X_n)])^2\right] \\ &= \mathbb{E}\left[g(X_1\ldots,X_n)^2\right]-(\mathbb{E}[g(X_1\ldots,X_n)])^2. \end{aligned} \tag{43}$$

*In practice, the variance is often analytically intractable, in which case, a linear approximation of the variance can be used. This involves expanding g about the means using a first-order Taylor series expansion around the point $= \{X_1 = \mathbb{E}[X_1]\ldots,X_n = \mathbb{E}[X_n]\}$*

$$g(X_1,\ldots,X_n) = g(point)+\sum_{i=1}^{n}(X_i-\mu_i)\frac{\partial g}{\partial X_i}\bigg|_{point}+\mathcal{O}(\partial^2 g) \tag{44}$$

*with this approximation, the variance of g can be approximated viz*

$$\begin{aligned} Var[g] &= \mathbb{E}\left[(g-\mathbb{E}[g])^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i-\mu_i)\frac{\partial g}{\partial X_i}\bigg|_{point}+\mathcal{O}(\partial^2 f)\right)^2\right] \\ &= \sum_{i=1}^{n}\frac{\partial g}{\partial X_i}\bigg|_{point}^2 Var[X_i]+\sum_{i\neq j}\frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}\bigg|_{point} Cov[X_i,X_j] \\ &\quad +\mathcal{O}(\partial^2 g). \end{aligned} \tag{45}$$

*where it has been used that $\mathbb{E}[g] = g(point)+\mathcal{O}(\partial^2 g)$ and it is understood all derivatives are evaluated at the means of the random variables. When the random variables are independent, $Cov[X_i,X_j] = 0$ for all $i \neq j$ (see theorem 12), and the formula simplifies to*

$$Var[g] \approx \sum_{i=1}^{n}\frac{\partial g}{\partial X_i}\bigg|_{point}^2 Var[X_i]. \tag{46}$$

**Example 3.6.**

*A company produce square plates. Take the dimensions of the plate to be characterized by two independent random variables* $X \sim \mathcal{N}(2m, (0.01m)^2)$ *and* $Y \sim \mathcal{N}(3m, (0.02m)^2)$ *and the area given by* $XY$*. Determine the variance of* $XY$*. From definition 35, the exact variance is*

$$\begin{aligned} Var[XY] &= \mathbb{E}[(XY)^2] - (\mathbb{E}[XY])^2 \\ &= \Big(Var[X] + \mathbb{E}[X]\Big)\Big(V[y] + \mathbb{E}[Y]\Big) - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\ &= \mathbb{E}[Y]^2 Var[X] + \mathbb{E}[X]^2 Var[Y] + Var[X]Var[Y] \end{aligned} \tag{47}$$

*where it has been used that* $X$ *and* $Y$ *are independent, such that* $\mathbb{E}[(XY)^2] = \mathbb{E}[X^2]\mathbb{E}[Y^2]$. *Via the linear approximation*

$$\begin{aligned} Var[XY] &\approx \sum_{i=X,Y} \left( \frac{\partial (XY)}{\partial i} \bigg|_{X=\mu_X, Y=\mu_Y} \right)^2 Var[i] \\ &= \mathbb{E}[Y]^2 Var[X] + \mathbb{E}[X]^2 Var[Y] \end{aligned} \tag{48}$$

*Comparing equation* (47) *and* (48) *the relative difference can be written*

$$\begin{aligned} \frac{Var[XY] - Var[XY]|_{\textit{linear approximation}}}{Var[XY]} &= \frac{Var[X]Var[Y]}{Var[XY]} \\ &\simeq 1.6 \cdot 10^{-5}. \end{aligned} \tag{49}$$

**Example 3.7.**

*Consider a thought experiment in which a father with amnesia is told he has two children, but does not know the sex of them. The sample space can be constructed from the sample space for each child*

$$\begin{aligned} \Omega_{\textit{child 1}} &= \{\text{🚹},\text{🚺}\}, \\ \Omega_{\textit{child2}} &= \{\text{🚹},\text{🚺}\} \end{aligned} \tag{50}$$

*such that*

$$\begin{aligned}\Omega &= \Omega_{child\ 1} \times \Omega_{child\ 2} \\ &= \{(\text{🚹},\text{🚹}), (\text{🚹},\text{🚺}), (\text{🚺},\text{🚹}), (\text{🚺},\text{🚺})\}.\end{aligned} \tag{51}$$

*Assuming the sex of a child is like a coin flip, it is most likely, a priori, that the father has one boy and one girl with probability $\frac{1}{2}$, i.e. $\mathbb{P}(\{(\text{🚹},\text{🚺})\}) = \frac{1}{2}$. The other possibilities (two boys or two girls) have probability $\frac{1}{4}$, meaning $\mathbb{P}(\{(\text{🚹},\text{🚹})\}) = \frac{1}{4}$ and $\mathbb{P}(\{(\text{🚺},\text{🚺})\}) = \frac{1}{4}$. In order to simplify the formalism, define the random variables $B : \Omega \mapsto \{0,1,2\}$ and $G : \Omega \mapsto \{0,1,2\}$ that maps the events in $\mathcal{F}$ to a number of boys $B(E) \forall E \in \mathcal{F}$ and girls $G(E) \forall E \in \mathcal{F}$. The probability mass function associated to $B$ and $G$ is given by equation* (27)*, such that e.g.*

$$p(B = 1, G = 1) = \mathbb{P}(\{(\text{🚹},\text{🚺})\}). \tag{52}$$

1. *Suppose the father ask his wife whether he has any boys, and she says yes. What is the probability that one child is a girl?*

   *The exact framing of the question is important here; "any boys" means "at least one boy"*

   $$p(G = 1, B \geq 1) = \frac{p(B \geq 1 | G = 1) p(G = 1)}{p(B \geq 1)}. \tag{53}$$

   *Given the father has two children, if he has exactly one girl, then the other must be a boy, so $p(B \geq 1 | G = 1) = 1$. $p(G = 1) = \frac{1}{2}$ since it is a priori assumed to be equally likely to be a boy or girl. $p(B \geq 1) = 1 - p(G = 2, B = 0) = \frac{3}{4}$, so*

   $$p(G = 1 | B \geq 1) = \frac{2}{3}. \tag{54}$$

2. *Suppose instead the father meets one of his children and it is a boy.What is the probability that the other is a girl?*

   *Since one child is known to be a boy, what is asked about is $p(G = 1 | B = 1) = \frac{1}{2}$.*

**Example 3.8.**

*Suppose a crime has been committed. Blood is found at the crime scene for which there is no innocent explanation. It is of the type which is present in* 1% *of the population.*

1. *The prosecutor claims: "There is a* 1% *chance that the defendant would have the crime blood type if he were innocent. Thus there is a* 99% *chance that he is guilty". This is known as the prosecutors fallacy. What is wrong with this argument?*

   *Let E denote the event of having the blood type found at the crime scene, then "there is a* 1% *chance that the defendant would have the crime blood type if he were innocent" means*

   $$\mathbb{P}(E|innocent) = 0.01. \tag{55}$$

   *This is not the relevant quantity, rather*

   $$\mathbb{P}(innocent|E) = \frac{\mathbb{P}(E|innocent)\mathbb{P}(innocent)}{p(E)}. \tag{56}$$

   *Since*

   $$\mathbb{P}(innocent|E) + \mathbb{P}(guilty|E) = 1 \tag{57}$$

   *and so* $\mathbb{P}(innocent|E) = 0.01$ *means* $\mathbb{P}(guilty|E) = 0.99$*, which is what is stated in the exercise, however, in general* $\mathbb{P}(E|innocent) \neq \mathbb{P}(innocent|E)$*.*

2. *The defender claims: "The crime occurred in a city of* 800 000 *people. Hence, the blood type found at the crime scene would be found in* $800\,000 \cdot 0.01 = 8\,000$ *people". The evidence has thus provided a probability of* $\frac{1}{8\,000}$ *that the defendant is guilty, and therefore has no relevance". This is known as the defendants fallacy. What is wrong with this argument?*

$$\mathbb{P}(guilty|E) = \frac{\mathbb{P}(E|guilty)\mathbb{P}(guilty)}{\mathbb{P}(E)}, \tag{58}$$

*with* $\mathbb{P}(E|guilty) = 1$, $\mathbb{P}(guilty) = \frac{1}{8\,000}$ *and*

$$\mathbb{P}(E) = \mathbb{P}(E|guilty)\mathbb{P}(guilty) + \mathbb{P}(E|innocent)p(innocent) \tag{59}$$

*where* $\mathbb{P}(E|innocent) = 0.01$ *and* $\mathbb{P}(innocent) = 1 - \mathbb{P}(guilty)$*, meaning*

$$\mathbb{P}(guilty|E) = \frac{100}{800\,099}. \tag{60}$$

$\frac{100}{800\,099}$ *is very close to* $\frac{1}{8\,000}$*, however, this assumes the only evidence against the defendant is the blood type found at the crime scene. If this changes, the calculation can change significantly, depending on the evidence.*

**Example 3.9.**

*Show that the variance of a sum is* $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$*.*

$$\begin{aligned} Var[X+Y] &= \mathbb{E}_{XY}[(X+Y-\mathbb{E}_{XY}[X+Y])^2] \\ &= \mathbb{E}_X[(X-\mathbb{E}_X[X])^2] + \mathbb{E}_Y[(Y-\mathbb{E}_Y[Y])^2] \\ &\quad + 2\mathbb{E}_{XY}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= Var[X] + Var[Y] + 2Cov[X,Y]. \end{aligned} \tag{61}$$

**Example 3.10.**

*After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 99%, as is the* *probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in* 10 000 *people. What are the chances that you actually have the disease?*

*Let "s" denote the event of being sick, "h" the event of being healthy, "p" the event of a positive test and "n" the event of a negative test, then*

$$\begin{aligned}\mathbb{P}(s|p) &= \frac{\mathbb{P}(p|sick)\mathbb{P}(s)}{\mathbb{P}(p)} \\ &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p|s)\mathbb{P}(s)+\mathbb{P}(p|h)\mathbb{P}(h)}\end{aligned} \tag{62}$$

*where* $\mathbb{P}(p|s) = 0.99$*,* $\mathbb{P}(s) = \frac{1}{10\,000}$*,* $\mathbb{P}(p|h) = 1 - \mathbb{P}(n|h)$*,* $\mathbb{P}(n|h) = 0.99$ *and* $\mathbb{P}(h) = 1 - \mathbb{P}(s)$*. This means*

$$\mathbb{P}(s|p) \simeq 0.0098. \tag{63}$$

---

**Example 3.11.**

*On a game show, a contestant is told the rules as follows: There are* 3 *doors labeled* 1, 2, 3*. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened, instead, the gameshow host will open one of the other two doors in such a way as not to reveal the prize. For example, if you first choose door* 1*, the gameshow host will open one of doors* 2 *and* 3*, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point you will be given a fresh choice of door: You can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.*
*Imagine that the contestant chooses first door* 1*; then the gameshow host opens door* 3*, revealing nothing. Should the contestant a) stick with door* 1*, b) switch to door* 2 *or c) it does not matter? You may assume that initially, the prize is equally likely to be behind*

*any of the* 3 *doors.*

*Let* $z_i$ *denote the prize being behind the i'th door,* $o_i$ *the action of opening the i'th door and* $c_i$ *the action of choosing the i'th door. The door with the largest probability of containing the prize should be picked, meaning*

$$z^* = \underset{z}{\operatorname{argmax}}(\mathbb{P}(z|o_3, c_1)). \tag{64}$$

*Since the host cannot open the door containing the prize,* $\mathbb{P}(z_3|o_3, c_1) = 0$ *and only* $\mathbb{P}(z_1|o_3, c_1)$ *and* $\mathbb{P}(z_2|o_3, c_1)$ *will have to be considered. For* $z_1$

$$\mathbb{P}(z_1|o_3, c_1) = \frac{\mathbb{P}(o_3|c_1, z_1)\mathbb{P}(c_1, z_1)}{\mathbb{P}(o_3, c_1)} \tag{65}$$

*with*

$$\begin{aligned} \mathbb{P}(o_3, c_1) &= \sum_i \mathbb{P}(o_3, c_1, z_i) \\ &= \mathbb{P}(o_3, c_1, z_1) + \mathbb{P}(o_3, c_1, z_2) + \mathbb{P}(o_3, c_1, z_3) \\ &= \mathbb{P}(o_3|c_1, z_1)\mathbb{P}(c_1, z_1) + \mathbb{P}(o_3|c_1, z_2)\mathbb{P}(c_1, z_2) \\ &\quad + \mathbb{P}(o_3|c_1, z_3)\mathbb{P}(c_1, z_3). \end{aligned} \tag{66}$$

$\mathbb{P}(o_3|c_1, z_3) = 0$ *since the host will not open the door with the prize.* $p(o_3|c_1, z_2) = 1$ *since the host has no other option in this case.* $\mathbb{P}(o_3|c_1, z_1) = \frac{1}{2}$ *since the host has two options in this case. There is no connection between the choice of door and position of the prize, so* $\mathbb{P}(c_1, z_j) = \mathbb{P}(c_1)\mathbb{P}(z_j)$ *and initially* $\mathbb{P}(z_j) = \mathbb{P}(z_k) \forall j, k \in \{1, 2, 3\}$. *Hence*

$$\begin{aligned} \mathbb{P}(z_1|o_3, c_1) &= \frac{\mathbb{P}(o_3|c_1, z_1)}{\sum_i \mathbb{P}(o_3|c_1, z_i)} \\ &= \frac{1}{3}. \end{aligned} \tag{67}$$

*Similarly*

$$\begin{aligned} \mathbb{P}(z_2|o_3, c_1) &= \frac{\mathbb{P}(o_3|c_1, z_2)}{\sum_i \mathbb{P}(o_3|c_1, z_i)} \\ &= \frac{2}{3}. \end{aligned} \tag{68}$$

*Since* $\mathbb{P}(z_2|o_3, c_1) > \mathbb{P}(z_1|o_3, c_1) > \mathbb{P}(z_3|o_3, c_1)$*, door number* 2 *is the optimal choise. Hence,answer "b)" is correct. The intuition behind the answer is the information the contestant has at the time of making the decision; initially, there is no a priori information and so* $\mathbb{P}(z_1|o_3, c_1) = \frac{1}{3}$*. At this time, there is* $\frac{2}{3}$ *probability that the prize is behind doors* 2, 3*. When the gameshow host open door* 3*, this probability converge on door* 2*.*

---

**Example 3.12.**

*Let* $X \sim Unif(a = -1, b = 1)$ *and* $Y = X^2$*. Clearly Y is dependent on X (in fact Y is uniquely determined by X). However, show that* $Corr[X, Y] = 0$*.*

$$\begin{aligned} Corr[X,Y] &= \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \\ &= \frac{\mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{Var[X]Var[Y]}} \end{aligned} \tag{69}$$

*In this case for the nonimator*

$$\begin{aligned} Cov[X,Y] &= \int dx x^3 p(X=x) - \int dx' x' p(X=x') \int dx'' x''^2 p(X=x'') \\ &= \frac{1}{b-a}\int_a^b x^3 dx - \frac{1}{(b-a)^2}\int_a^b dx' x' \int_a^b dx'' x''^2 \\ &= \frac{1}{12}(a-b)^2(a+b) \\ &= 0 \end{aligned} \tag{70}$$

*where the last equality comes from the fact that* $a + b = 0$ *in this case. However, we need to make sure the denominator does not diverge*

$$\begin{aligned} Var[X]Var[X^2] &= \left(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2\right)\left(\mathbb{E}_X[X^4] - \mathbb{E}_X[X^2]^2\right) \\ &= \frac{1}{540}(b-a)^4(4a^2 + 7ab + 4b^2) \\ &\neq 0. \end{aligned}$$

$$\tag{71}$$

*It denominator does not diverge, so the factorized $a + b$ from the nominator makes $Corr[X, X^2] = 0$.*

---

**Example 3.13.**

*Let $X \sim N(\mu = 0, \sigma^2 = 1)$ and $Y = WX$, where $W$ is a discrete random variable defined by $p(W = -1) = p(W = 1) = \frac{1}{2}$. It is clear that $X$ and $Y$ are not independent, since $Y$ is a function of $X$.*

1. *Show $Y \sim N(\mu = 0, \sigma^2 = 1)$.*

   *To show that $Y \sim N(\mu = 0, \sigma^2 = 1)$, show that $Y$ has zero mean and unity variance.*

$$\begin{aligned} \mathbb{E}_Y[Y] &= \mathbb{E}_{WX}[WX] \\ &= \mathbb{E}_W[W]\cancelto{0}{\mathbb{E}_X[X]} \\ &= 0. \end{aligned} \tag{72}$$

*The variance*

$$\begin{aligned} Var[Y] &= \mathbb{E}_Y[Y^2] - \cancelto{0}{\mathbb{E}_Y[Y]^2} \\ &= \mathbb{E}_{WX}[W^2X^2] \\ &= \mathbb{E}_W[W^2]\mathbb{E}_X[X^2] \\ &= \mathbb{E}_W[W^2]Var[X] \end{aligned} \tag{73}$$

*since $Var[X] = \mathbb{E}_X[X^2] - \cancelto{0}{\mathbb{E}_X[X]^2} = 1$. Now*

$$\begin{aligned} \mathbb{E}_W[W^2] &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 p(W = w_i) \\ &= \frac{1}{2}[(-1)^2\frac{1}{2} + 1^2\frac{1}{2}] \\ &= 1 \end{aligned} \tag{74}$$

*so $Var[Y] = 1$.*

2. *Show $Cov[X, Y] = 0$. Thus X and Y are uncorrelated but dependent, even though they are Gaussian.*

$$\begin{aligned} Cov[X, Y] &= Cov[X, WX] \\ &= \mathbb{E}_{WX}[WX^2] - \mathbb{E}_X[X]\mathbb{E}_{WX}[WX] \\ &= \mathbb{E}_W[W]\mathbb{E}_X[X^2] - \mathbb{E}_W[W]\mathbb{E}_X[X]^2 \\ &= \mathbb{E}_W[W]Var[X] \\ &= 0 \end{aligned} \tag{75}$$

*where for the last equality it has been used that*

$$\begin{aligned} \mathbb{E}_W[W] &= \frac{1}{n}\sum_{i=1}^{n} w_i p(W = w_i) \\ &= \frac{1}{2}[(-1)\frac{1}{2} + 1\frac{1}{2}] \\ &= 0 \end{aligned} \tag{76}$$

**Example 3.14.**

*Prove that $-1 \leq Corr[X, Y] \leq 1$.*

*Since the variance is defined as positive definite*

$$\begin{aligned} 0 \leq & Var\left[\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right] \\ &= \frac{Var[X]}{\sigma_X^2} + \frac{Var[Y]}{\sigma_Y^2} \pm \frac{2}{\sigma_X\sigma_Y}Cov[X, Y] \\ &= \frac{Var[X]}{\sigma_X^2} + \frac{Var[Y]}{\sigma_Y^2} \pm 2Corr[X, Y] \\ &= 2 \pm 2Corr[X, Y] \end{aligned} \tag{77}$$

*where for the last equality it has been used that $\sigma_i^2 = Var[i]$.* $0 \leq 2 \pm 2Corr[X, Y] \Leftrightarrow -1 \leq Corr[X, Y] \leq 1$.

**Example 3.15.**

*Show that if $Y = aX + b$ for some parameters $a > 0$ and $b$, then $Corr[X, Y] = 1$. Similarly show that if $a < 0$, then $Corr[X, Y] = -1$.*

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}} \tag{78}$$

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\ &= \mathbb{E}_X[X(aX + b)] - \mathbb{E}_X[X]\mathbb{E}_X[aX + b] \\ &= a\mathbb{E}_X[X^2] + b\mathbb{E}_X[X] - a\mathbb{E}_X[X]^2 - b\mathbb{E}_X[X] \\ &= aVar[X] \end{aligned} \tag{79}$$

$$\begin{aligned} Var[Y] &= Var[aX + b] \\ &= a^2Var[X] + \cancelto{0}{Var[b]} + 2\cancelto{0}{Cov[aX, b]} \\ &= a^2Var[X] \end{aligned} \tag{80}$$

$$\begin{aligned} Corr[X, Y] &= \frac{aVar[X]}{\sqrt{a^2Var[X]Var[X]}} \\ &= \frac{a}{|a|} \end{aligned} \tag{81}$$

*Hence, the sign of "a" determine if $Corr[X, Y] = \pm 1$ for the particular $Y$ of this example.*

# CHAPTER 4

# Introduction to Statistics

Let the observed outcome of a statistical experiment be described by the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (see chapter 3), where as opposed to the case in probability theory, $\mathbb{P}$ is now unknown. A generic number of random variables are defined on the sample space viz [4, 7–9]

$$X_i : \Omega \mapsto \Omega_{X_i}, \tag{82}$$

where $\Omega_{X_i}$ is part of the probability space $(\Omega_{X_i}, \mathcal{F}_{X_i}, \mathbb{P}_{X_i})$, where

$$\mathbb{P}_{X_i} = \mathbb{P} \circ X_i^{-1} \tag{83}$$

is the push forward measure (see definition 29) of $\mathbb{P}$ with respect to $X_i$. The joint probability measure can be defined viz

$$\mathbb{P}_{X_1,\ldots X_n} = \mathbb{P} \circ (X_1, \ldots X_n)^{-1}. \tag{84}$$

on the measurable space

$$(\Omega_{X_1} \cdots \times \Omega_{X_n}, \mathcal{F}_{X_1} \cdots \otimes \mathcal{F}_{X_n}) \tag{85}$$

which for brevity will be written $(\Omega_{X_{1:n}}, \mathcal{F}_{X_{1:n}})$. Depending on the discrete or continuous nature of the different random variables, there are discrete (PMF, see definition 30) or continuous probability distributions (PDF, see definition 31) associated to the joint probability measure. All probability distributions related to the random variables can be derived from the joint probability distribution via marginalization (see theorem 3).

**Definition 36** (Set of Probability Measures). *Let $\mathcal{P}$ be the set of all probability measures on $(\Omega_{X_{1:n}}, \mathcal{F}_{X_{1:n}})$. It is assumed, often based on prior information, that $\mathbb{P}_{X_1,\ldots X_n} \in \mathcal{P}' \subseteq \mathcal{P}$, which is described in parametric form viz*

$$\mathcal{P}' = \{\mathbb{P}_{X_1,\ldots X_n}(w) | w \in \Omega_W\}, \tag{86}$$

*where $\Omega_W$ is called the parameter space.*

**Definition 37** (Parameter Space). *$\mathbb{P}_{X_1,\ldots X_n}(w) \in \mathcal{P}'$ is specified by parameters $w \in \Omega_W$, where $\Omega_W$ is the parameter space.*

**Definition 38** (Identifiable statistical model). *A statistical model is identifiable if $w \in \Omega_W \mapsto \mathbb{P}_{X_1,\ldots X_n}(w) \in \mathcal{P}'$ is injective (one-to-one).*

The parameters $w \in \Omega_W$ can either be viewed as fixed constants or the realization of a random variable.

**Axiom 4** (Parameter Fixedness). *The parameter $w \in \Omega_W$ is treated as a fixed but unknown constant in the statistical model.*

**Axiom 5** (Parameter as a Random Variable). *The parameter $w \in \Omega_W$ is treated as a realization of a random variable. In this case, the parameter space must be endowed with a $\sigma$-algebra ($\mathcal{F}_W$) and a probability measure ($\mathbb{P}_W$) that must be the result of another measure pushed forward (see definition 29) with respect to the random variable $W$. This means*

$$W : \Omega \mapsto \Omega_W \tag{87}$$

*is defined as a random variable that maps from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the probability space $(\Omega_W, \mathcal{F}_W, \mathbb{P}_W)$, and where*

$$\mathbb{P}_W : \mathcal{F}_W \mapsto [0, 1], \tag{88}$$

*is called the prior measure, which is the push forward measure of $\mathbb{P}$ with respect to $W$, i.e.*

$$\mathbb{P}_W = \mathbb{P} \circ W^{-1}. \tag{89}$$

For both Axiom 4 and Axiom 5, the value of a parameter is considered fixed. Axiom 5 introduces a random variable $W$ not to add randomness to the parameter $w$ but to model uncertainty or variability about the fixed but unknown parameter value. Observations of the random variables $X_1, \dots X_n$ are used to a) estimate the parameters if they are fixed and b) estimate the joint probability distribution of the parameters if they are random variables. Hence, given a set of observations of the random variables $X_1, \dots X_n$ and defining an appropriate subset $\mathcal{P}'$ for the joint probability measure, probability theory can be used to answer statistical questions. This highlights the dual nature of statistics, comprised of two integral parts.

1. The first part involves the formulation and evaluation of probabilistic models, a process situated within the realm of the philosophy of science. This phase grapples with the foundational aspects of constructing models that accurately represent the problem at hand.

2. The second part concerns itself with extracting answers after assuming a specific model. Here, statistics becomes a practical application of probability theory, involving not only theoretical considerations but also numerical analysis in real-world scenarios.

This duality underscores the interdisciplinary nature of statistics, bridging the gap between the conceptual and the applied aspects of probability theory.

## 4.1 INTERPRETATION OF A PROBABILITY MEASURE

Although probability measures are well defined (see chapter 3), their interpretation is not defined beyond their definition. For this reason there are two broadly accepted interpretations of probability; objective and subjective.

**Definition 39** (Objective Probability Measure). *Let $\mathbb{P}$ denote a generic probability measure defined on the generic probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The "objective probability measure"-interpretation define $\mathbb{P}$ as the long-run or limiting frequency of an event, $E$. That is, let $m$ be the number of occurrences of $E$, and let $n$ be the number of experiments, then [10]*

$$\mathbb{P}(E) \equiv \lim_{n \to \infty} \left( \frac{m}{n} \right) \tag{90}$$

*define the probability measure as the limit of a relative frequency.*

**Definition 40** (Sugeno Measure). *Let $(\Omega, \mathcal{F})$ be a measurable space (definition 22) and $Bel : \mathcal{F} \to [0, 1]$ a Sugeno measure iff [11]*

1. ***Non-negativity**: $Bel(\emptyset) = 0$,*
2. ***Normalization**: $Bel(\Omega) = 1$,*
3. ***Monotonicity**: For all $A, B \in \mathcal{F}$, if $A \subseteq B$, then $Bel(A) \leq Bel(B)$.*

**Definition 41** (Subjective Probability Measure). *A subjective probability measure is a numerical representation of rational beliefs. Formally, it is a probability measure (definition 23) $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{F})$ that fulfills the definition of a Sugeno measure (definition 40) [11, 12].*

**Theorem 13.** *Any probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a Sugeno measure.*

*Proof.* Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$. By definition, $\mathbb{P}$ satisfies:

1. $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$ (Boundary Conditions).
2. If $A, B \in \mathcal{F}$ and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$ (Monotonicity).

Thus, $\mathbb{P}$ is a Sugeno measure. □

**Corollary 1.** *Since a probability measure $\mathbb{P}$ satisfies the axioms of a Sugeno measure, it can be interpreted as a belief function.*

**Definition 42** (Frequentist Statistics). *Frequentist statistics is a paradigm that adopts axiom 4 and the definition 39 of probability.*

**Definition 43** (Bayesian Statistics). *Bayesian statistics is a paradigm that adopts axiom 5 and the definition 41 of probability.*

**Example 4.1.**

*In the Frequentist approach one can say; if an experiment is repeated many times, in (e.g.) 95% of these cases the computed confidence interval will contain the true parameter value.*
*In the Bayesian approach one can say; given the observed data, there is a 95% probability that the value of the true parameter lies within the Bayesian interval.*
*Note how in the Frequentist approach the true parameter is fixed and the confidence interval is varying. In the Bayesian approach the interval is fixed and the true parameter is varying.*

**Example 4.2.**

*Consider a Bayesian statistical model involving both a normal distribution and a beta distribution. The parameter space $\mathbb{W}$ includes the parameters $\mu$ and $\sigma$ representing the mean and standard deviation, respectively, for the normal distribution, as well as the parameters $a$ and $b$ for the beta distribution, meaning*

$$\mathbb{W} = \{\mu, \sigma, a, b\}. \tag{91}$$

*The random variable $W : \Gamma \mapsto \mathbb{W}$ must be a vector*

$$W = \begin{pmatrix} W_\mu & W_\sigma & W_a & W_b \end{pmatrix}^T, \tag{92}$$

*such that each individual parameter has an associated probability distribution.*

## 4.2 RELAXATION OF NOTATION

Fortunately, a lot of the details around probability spaces and measures can be abstracted in the practical application of statistics. For this reason, in the remainder of the book, where the practical application of statistics is considered, the notation and formalization especially around probability spaces, algebras, probability measures ect. is relaxed considerably – which is the norm, by the way. Specifically, in the rest of this book, $p$ will be used to denote anything related to probability distributions or measures and the probability for a random variable to take on a specific value , e.g. $p(X = x)$, will usually be denoted $p(x)$ for shorthand. This relaxation of notation facilitates advanced manipulation of probabilities, which would otherwise be incredibly cumbersome. It is, however, beneficial to have some background knowledge about the formal definitions, hence this introduction.

CHAPTER 5

# Assigning Probability Functions

The axioms and definitions of probability theory (axioms 1-3 and definitions 8 and 9) can be used to define and relate probability measures, however, they are not sufficient to conduct inference because, ultimately, the probability measure or relevant probability functions (density or mass) needs to be specified. Thus, the rules for manipulating probability functions must be supplemented by rules for assigning probability functions. To assign any probability function, there is ultimately only one way, logical analysis, i.e., non-self-contradictory analysis of the available information. The difficulty is to incorporate only the information one actually possesses without making gratuitous assumptions about things one does not know. A number of procedures have been developed that accomplish this task: Logical analysis may be applied directly to the sum and product rules to yield probability functions [13]. Logical analysis may be used to exploit the group invariances of a problem [14]. Logical analysis may be used to ensure consistency when uninteresting or nuisance parameter are marginalized from probability functions [15]. And last, logical analysis may be applied in the form of the principle of maximum entropy to yield probability functions [14, 16–19]. Of these techniques the principle of maximum entropy is probably the most powerful.

## 5.1 THE PRINCIPLE OF MAXIMUM ENTROPY

The principle of maximum entropy, first proposed by Jaynes [20], considers the issue of assigning a probaility distribution to a random variable. Let $Z$ be a generic random variable that describes an abstract experiment. $Z$ follow a distribution $p(z|\lambda, I)$ with associated parameters $\lambda = \{\lambda_0, \ldots, \lambda_n\}$. The principle of maximum entropy propose that the probability distribution, $p(z|\lambda, I)$, which best represents the current state of knowledge about a system is the one with largest constrained entropy [1], defined by the Lagrangian

$$\mathcal{L} = \int F dz, \tag{93}$$

with

$$F = -p(z|\lambda, I) \ln \frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \sum_{j=1}^{n} \lambda_j C_j(z). \tag{94}$$

$m$ – called the Lebesgue measure – ensures the entropy, given by $-\int p(z|\lambda, I) \ln \frac{p(z|\lambda,I)}{m(z)} dz$, is invariant under a change of variables and $C_j(z)$ represent the constraints beoynd normalization. The constraint beyond normality depend on the background information related to the random variable, $X$. In variational calculus the Lagrangian is optimized via solving the Euler-Lagrange equation

$$\frac{\partial F}{\partial p(z|\lambda, I)} - \frac{d}{dx} \frac{\partial F}{\partial p(z|\lambda, I)'} = 0, \tag{95}$$

where $\frac{\partial p(z|\lambda,I)}{\partial x} = p(z|\lambda, I)'$ for shorthand. Since $p(z|\lambda, I)' \notin F$, the Euler-Lagrange equation simplify to simply

$$\frac{\partial F}{\partial p(z|\lambda, I)} = 0. \tag{96}$$

Combining equations (93) and (96)

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -\ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - 1 - \sum_j \lambda_j C_j(z) \\ &= 0 \end{aligned} \tag{97}$$

and so

$$\begin{aligned} p(z|\lambda, I) &= m(z)e^{-1-\sum_j \lambda_j C_j(z)} \\ &= \tilde{m}(z)e^{-\sum_j \lambda_j C_j(z)}, \end{aligned} \tag{98}$$

where $\tilde{m}(z) \equiv m(z)e^{-1}$. Using that $\int p(z|\lambda, I)dx = 1$

$$p(z|\lambda, I) = \frac{\tilde{m}(z)e^{-\sum_j \lambda_j C_j(z)}}{\int \tilde{m}(z')e^{-\sum_j \lambda_j C_j(z')}dz'}, \tag{99}$$

where $m$ is a reference distribution that is invariant under parameter transformations. $\lambda_j$ are determined from the additional constraints, e.g. on the mean or variance.

**Example 5.1.**

*Consider a random variable, Z, with unlimited support, $z \in [-\infty, \infty]$, assumed to be symmetric around a single peak defined by the mean $\mu$, standard deviation $\sigma$. In this case $\lambda = \{\lambda_0, \lambda_1, \lambda_2\}$, where it will be shown that $\lambda_1, \lambda_2$ are related to $\mu, \sigma$. In this case F can be written*

$$\begin{aligned} F = &- p(z|\lambda, I)\ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_0 p(z|\lambda, I) \\ &- \lambda_1 p(z|\lambda, I)z - \lambda_2 p(z|\lambda, I)z^2 \end{aligned} \tag{100}$$

*with the derivative*

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_1 z - \lambda_2 z^2 \\ &= 0, \end{aligned} \tag{101}$$

*meaning*

$$p(z|\lambda, I) = m(z)e^{-1-\lambda_0-\lambda_1 z-\lambda_2 z^2}. \tag{102}$$

*Taking a unifoirm measure ($m = const$) and imposing the normalization constraint*

$$\begin{aligned} \int p(z|\lambda, I) dz &= m e^{-1-\lambda_0} \int e^{-\lambda_1 z - \lambda_2 z^2} dz \\ &= m e^{-1-\lambda_0} \sqrt{\frac{\pi}{\lambda_2}} e^{\frac{\lambda_1^2}{4\lambda_2}} \\ &= 1. \end{aligned} \tag{103}$$

*Defining $K^{-1} = m e^{-1-\lambda_0}$ yields*

$$\begin{aligned} p(z|\lambda, I) &= \frac{e^{-\lambda_1 x - \lambda_2 x^2}}{K} \\ &= \sqrt{\frac{\lambda_2}{\pi}} e^{-\frac{\lambda_1^2}{4\lambda_2} - \lambda_1 z - \lambda_2 z^2} \end{aligned} . \tag{104}$$

*Now, imposing the mean constraint*

$$\begin{aligned} \int z p(z|\lambda, I) dz &= \frac{\int z e^{-\lambda_1 z - \lambda_2 z^2} dz}{K} \\ &= -\frac{\lambda_1}{2\lambda_2} \\ &= \mu. \end{aligned} \tag{105}$$

*Hereby*

$$\begin{aligned} p(z|\lambda, I) &= \sqrt{\frac{\lambda_2}{\pi}} e^{-\mu^2 \lambda_2 + 2\mu\lambda_2 z - \lambda_2 z^2} \\ &= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{\mu - z}{\sigma}\right)^2}, \end{aligned} \tag{106}$$

*where $\sigma \equiv \frac{1}{2\lambda_2}$ has been defined. Hence, it is clear that the normal distribution can be derived from general constraints via the principle of maximum entropy.*

**Example 5.2.**

*Consider a random variable, Z, with limited support,* $z \in [0, 1]$. *In order to impose the limited support, require that* $\ln(z)$ *and* $\ln(1-z)$ *be well defined. In this case F can be written*

$$\begin{aligned} F = &- p(z|\lambda, I)\ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_0 p(z|\lambda, I) \\ &- \lambda_1 p(z|\lambda, I)\ln(z) - \lambda_2 p(z|\lambda, I)\ln(1-z) \end{aligned} \tag{107}$$

*with the derivative*

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_1 \ln(z) - \lambda_2 \ln(1-z) \\ &= 0, \end{aligned} \tag{108}$$

*meaning*

$$p(z|\lambda, I) = m(z)e^{-1-\lambda_0-\lambda_1\ln(z)-\lambda_2\ln(1-z)}. \tag{109}$$

*Taking a unifoirm measure (m = const) and imposing the normalization constraint*

$$\begin{aligned} \int p(z|\lambda, I)dz &= me^{-1-\lambda_0}\int z^{-\lambda_1}(1-z)^{-\lambda_2}dz \\ &= me^{-1-\lambda_0}\frac{\Gamma(1-\lambda_1)\Gamma(1-\lambda_2)}{\Gamma(2-\lambda_1-\lambda_2)} \\ &= 1. \end{aligned} \tag{110}$$

*Now define* $\alpha \equiv 1-\lambda_1 \wedge \beta \equiv 1-\lambda_2$. *Hereby*

$$p(z|\alpha, \beta, I) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{111}$$

*which is the beta distribution.*

# CHAPTER 6

## Framing of Statistics

In this book, the field of statistics will be framed as a game against Nature, as is conventionally done i decision theory. In this game there are two players or decision makers

1. **Robot:** This is the name given to the primary decision maker.

2. **Nature:** This decision maker is a mysterious entity that is unpredictable to the Robot. It has its own set of actions, and it can choose them in a way that interferes with the achievements of the Robot. Nature can be considered as a synthetic decision maker that is constructed for the purposes of modeling uncertainty in the decision-making or planning process.

The game is described by the interaction between the Robot and Nature, characterized by the probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, the parameter space $\Omega_W$, and the set of probability distributions $\mathcal{P}$ parameterized by the parameters $w \in \Omega_W$. Imagine that the Robot and Nature each make a decision by choosing an action from a set, $u \in \Omega_U$ and $s \in \Omega_S$, respectively. $\Omega_U$ is referred to as the action space, and $\Omega_S$ as the Nature action space. The Robot receives a numerical penalty, assigned by a cost function, depending on the two decisions made.

**Definition 44** (Cost Function)**.** *A cost function associates a numerical penalty depending on decision $u \in \Omega_U$ and $s \in \Omega_S$,*

$$C : \Omega_U \times \Omega_S \mapsto \mathbb{R}. \tag{112}$$

Given the observation $X = x$ as well as a set of past observations and matching actions of Nature $D = \{(x_i, s_i)|i = 1 : n\}$, the Robot's objective is to formulate a decision rule that minimize the expected cost associated with its decisions.

**Definition 45** (Decision Rule). *A decision rule is a function $U$ that maps from the observation space $\Omega_X$ and past observations and decisions $\Omega_X^n \times \Omega_S^n$ to a set of possible actions $\Omega_U$, meaning*

$$U : \Omega_X \times \Omega_S \mapsto \Omega_U. \tag{113}$$

**Example 6.1.**

*Suppose the Robot has an umbrella and considers if it should bring it on a trip outside, i.e.*

$$\mathbb{U} = \{\text{"bring umbrella"}, \text{"don't bring umbrella"}\}. \tag{114}$$

*Nature have already picked whether or not it will rain later, i.e.*

$$\Omega_S = \{\text{"rain"}, \text{"no rain"}\}, \tag{115}$$

*so the Robot's task is to estimate Nature's decision regarding rain later and either bring the umbrella or not. The Robot's decision rule, denoted as $U$, maps the available information $X = x$ (possibly $X =$weather forecasts, current weather conditions, etc.) to one of its possible actions. For instance, $U$(weather forecast) might map to the action "bring umbrella" if rain is predicted and "don't bring umbrella" otherwise.*

---

The random variable $X : \Omega \mapsto \Omega_X$ represent the information available (the information may be missing or null) to the Robot regarding the decision Nature will make, while $S : \Omega \mapsto \Omega_S$ represent the different possible decisions of Nature. $\Omega_X$ and $\Omega_S$ have associated $\sigma$-algebras and probability measures, however, such details are assumed *to be understood* in the practical application of statistics. Given the observation $X = x$ as well as a set of past observations

$$D = \{(X = x_1, S = s_1), \dots (X = x_n, S = s_n)\}, \tag{116}$$

the objective of the Robot is to minimize the expected cost associated with its decisions [2]

$$\begin{aligned}\mathbb{E}[C(U,S)|I] &= \int dDdxdsC(U(x,D),s)p(X=x,S=s,D|I)\\ &= \int d\tilde{D}dsC(U(\tilde{D}),s)p(S=s,\tilde{D}|I)\end{aligned} \tag{117}$$

where $\tilde{D} = \{D, X = x\}$ and the Robot aims to find the decision rule which minimizes equation (163), meaning

$$U^* = \arg\min_U \mathbb{E}[C(U,S)|I]. \tag{118}$$

From theorem 10

$$\mathbb{E}[C(U,S)|I] = \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]]. \tag{119}$$

Using equation (165) in equation (164)

$$\begin{aligned}U^* &= \arg\min_U \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]]\\ &= \arg\min_U \int dxp(\tilde{D}|I)\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I].\end{aligned} \tag{120}$$

Since $p(\tilde{D}|I)$ is a non-negative function, the minimizer of the integral is the same as the minimizer of the conditional expectation, meaning

$$\begin{aligned}U^*(\tilde{D}) &= \arg\min_{U(\tilde{D})} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]\\ &= \arg\min_{U(\tilde{D})} \int dsC(U(\tilde{D}),s)p(S=s|X=x,D,I).\end{aligned} \tag{121}$$

**Example 6.2.**

*In general the random variable X represent the observations the Robot has available that are related to the decision Nature is going* *to make. However, this information may not be given, in which case* $\{x, D_x\} = \emptyset$ *and consequently*

$$\begin{aligned}\tilde{D} &= \{S_1 = s_1, \dots S_n = s_n\} \\ &\equiv D_s.\end{aligned} \tag{122}$$

*In this case, the Robot is forced to model the decisions of Nature with a probability distribution with associated parameters without observations. From equation* (274) *the optimal action for the Robot can be written*

$$U^*(D_s) = \arg\min_{U(D_s)} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] \tag{123}$$

---

## 6.1 ASSIGNING A COST FUNCTION

The cost function (see definition 44) associates a numerical penalty to the Robot's action and thus the details of it determine the decisions made by the Robot. Under certain conditions, a cost function can be shown to exist [3], however, there is no systematic way of producing or deriving the cost function beyond applied logic. In general, the topic can be split into considering a continuous and discrete action space, $\Omega_U$.

### 6.1.1 *Continuous Action Space*

In case of a continuous action space, the cost function is typically picked from a set of standard choices.

**Definition 46** (Linear Cost Function)**.** *The linear cost function is defined viz*

$$C(U(\tilde{D}), s) \equiv |U(\tilde{D}) - s|. \tag{124}$$

**Theorem 14** (Median Decision Rule). *Assuming the cost function of definition 52*

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] &= \int_{-\infty}^{\infty} ds|U(\tilde{D})-s|p(s|\tilde{D},I)\\ &= \int_{-\infty}^{U(\tilde{D})}(s-U(\tilde{D}))p(s|\tilde{D},I)ds\\ &\quad+\int_{U(\tilde{D})}^{\infty}(U(\tilde{D})-s)p(s|\tilde{D},I)ds\end{aligned} \tag{125}$$

$$\begin{aligned}0 &= \left.\frac{d\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]}{dU(\tilde{D})}\right|_{U(\tilde{D})=U^*(\tilde{D})}\\ &= (U^*(\tilde{D})-U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D},I)+\int_{-\infty}^{U^*(\tilde{D})}p(s|\tilde{D},I)ds\\ &\quad+(U^*(\tilde{D})-U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D},I)-\int_{U^*(\tilde{D})}^{\infty}p(s|\tilde{D},I)ds\end{aligned} \tag{126}$$

$$\begin{aligned}\int_{-\infty}^{U^*(\tilde{D})}p(s|\tilde{D},I)ds &= \int_{U^*(\tilde{D})}^{\infty}p(s|\tilde{D},I)ds\\ &= 1-\int_{-\infty}^{U^*(\tilde{D})}p(s|\tilde{D},I)ds\end{aligned} \tag{127}$$

$$\int_{-\infty}^{U^*(\tilde{D})}p(s|\tilde{D},I)ds = \frac{1}{2} \tag{128}$$

*which is the definition of the median.*

**Definition 47** (Quadratic Cost Function). *The quadratic cost function is defined as*

$$C(U(\tilde{D}),s) \equiv (U(\tilde{D})-s)^2. \tag{129}$$

**Theorem 15** (Expectation Decision Rule). *Assuming the cost function of definition 53*

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] &= \int ds(U(\tilde{D})-s)^2 p(s|\tilde{D},I)\\ &\Downarrow\\ \frac{d\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]}{dU(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(x)} &= 2U^*(\tilde{D})-2\int dssp(s|\tilde{D},I)\\ &= 0\\ &\Downarrow\\ U^*(\tilde{D}) &= \int dssp(s|\tilde{D},I)\\ &= \mathbb{E}[S|\tilde{D},I]\end{aligned} \tag{130}$$

*which is the definition of the expectation value.*

**Definition 48** (0-1 Cost Function). *The 0-1 cost function is defined viz*

$$C(U(\tilde{D}),s) \equiv 1-\delta(U(\tilde{D})-s). \tag{131}$$

**Theorem 16** (MAP Decision Rule). *The maximum aposteriori (MAP) follows from assuming 0-1 loss viz*

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C((\tilde{D}),S)|\tilde{D},I] &= 1-\int ds\delta(U(\tilde{D})-s)p(s|\tilde{D},I)\\ &\Downarrow\\ \frac{d\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]}{dU(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(\tilde{D})} &= -\frac{dp(s|\tilde{D},I)}{ds}\bigg|_{s=U^*(\tilde{D})}\\ &= 0\end{aligned} \tag{132}$$

*which is the definition of the MAP.*

### 6.1.2 *Discrete Action Space*

In case of a continuous action space, the conditional expected loss can be written

$$\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] = \sum_{s\in\mathbb{S}}^{n} C(U(\tilde{D}),s)p(s|\tilde{D},I), \tag{133}$$

where the cost function is typically represented in matrix form viz

$$\begin{array}{cc|ccc|}
\multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{3}{c}{S} \\
\multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{s_1} & \multicolumn{1}{c}{\dots} & \multicolumn{1}{c}{s_{\dim(\Omega_S)}} \\ \cline{3-5}
U(x) & u_1 & C(u_1,s_1) & \dots & C(u_1,s_{\dim(\Omega_S)}) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & u_{\dim(\Omega_U)} & C(u_{\dim(\Omega_U)},s_1) & \dots & C(u_{\dim(\Omega_U)},s_{\dim(\Omega_S)}) \\ \cline{3-5}
\end{array}$$

## 6.2 STATISTICAL PARADIGMS

So far in this chapter, there has been no reference to the statistical paradigms (Bayesian and Frequentist). This is because all so far is valid for both the Bayesian (definition 43) and Frequentist (definition 4.1) paradigms. The difference between the two comes to light when considering the parameters of Natures model.

# Part II

# FREQUENTIST STATISTICS

# CHAPTER 7

## Frequentist Statistics Introduction

Frequentist statistics is based on definition 4.1, which follows the definition of objective probability (definition 39) and the principle of fixed, unknown parameters (axiom 4). The foundations of Frequentist statistics trace back to seminal works such as those of Neyman and Pearson [21] and Fisher [22], who laid the groundwork for much of its methodology. Subsequent developments by Wald [23], Neyman [24], and Lehmann [25] further refined its theories and techniques.

In the Frequentist paradigm, it is assumed that Natures decisions can be captured by a model with unknown, fixed parameters $w$. This means everything in chapter 6 becomes conditioned on $w$, and the focus becomes estimating $w$ via an estimator $\hat{w}(\tilde{D})$ and subsequently deciding the optimal decision rule

$$U^*(X = x, \hat{w}(\tilde{D})) \tag{134}$$

according to a cost function, as specified in chapter 6.

**Example 7.1.**

*Let $X$ and $S$ be continuous random variables with the relationship [26]*

$$S = f(X, w) + \epsilon \tag{135}$$

*where $\epsilon$ is a random variable representing noise, with $\mathbb{E}[\epsilon] = 0$ and $f$ is some model with parameters $w$. Using the quadratic cost function of definition 53 theorem* (21) *yields*

$$\begin{aligned} U^*(X = x, w) &= \mathbb{E}[S|X = x, w, I] \\ &= f(x, w), \end{aligned} \tag{136}$$

*where it has been used that $f$ does not depend on past data. Hence, if $w$ was known, the Robot could map any observation $X = x$ using the optimal decision rule 136. In reality $w$ is not known and must be estimated.*

# CHAPTER 8

## Parameter Estimation

Given a decision rule, $U(x, w)$, the unknown parameter $w$ need to be estimated. This is unique to frequentist statistics as the parameters $w$ are considered realizations of random variables in Bayesian statistics. $w$ is estimated by reapplying decision theory. To distinguish this scenario from previous ones, denote the decision rule in this case $\hat{w}$, then

**Definition 49** (Fisher Information)**.** *The Fisher information is a way of measuring the amount of information about an unknown parameter a random variable contains. Let $w$ be an unknown parameter, $p(X|w)$ a probability distribution for the generic random variable $X : \Omega \mapsto \Omega_X$ and define $l(X|w) = \frac{\partial}{\partial w} \ln p(X|w)$. The Fisher information can then be written*

$$\begin{aligned} I(w) &\equiv \mathbb{E}[l(X|w)^2|w] \\ &= Var[l(X|w)|w]. \end{aligned} \tag{137}$$

*Proof.* In general

$$\mathbb{E}[l(X|w)^2|w] = \text{Var}[l(X|w)|w] + \mathbb{E}[l(X|w)|w]^2 \tag{138}$$

however

$$\begin{aligned} \mathbb{E}[l(X|w)|w] &= \int \frac{\partial}{\partial w} \ln p(X = x|w) p(X = x|w) dx \\ &= \frac{\partial}{\partial w} \int p(X = x|w) dx \\ &= 0. \end{aligned} \tag{139}$$

□

**Theorem 17** (Fisher information for sample). *Let $X_1, X_2, \dots X_n$ be a set of independent and identically distributed random variables from the measurable space $(\Omega, \mathcal{F})$. The Fisher information in a sample is*

$$I(w) = nI_1(w), \tag{140}$$

*where $I_1(w)$ is the Fisher information of any one of the random variables.*

**Definition 50** (Maximum Likelihood Estimator (MLE) Decision Rule). *The Maximum Likelihood Estimator (MLE) decision rule $\hat{w}_{MLE}$ is defined as the decision rule that maximizes the likelihood $p(D_s|D_x, w)$ given the data $D_x$ and past Nature decisions $D_s$*

$$\hat{w}_{MLE}(\tilde{D}) \equiv \arg\max_{w} p(D_s|D_x, w). \tag{141}$$

**Theorem 18** (Unbiasedness of the MLE Decision Rule). *Under certain regularity conditions, the MLE decision rule $\hat{w}_{MLE}$ is asymptotically unbiased, meaning*

$$\sqrt{n}(\hat{w}_{MLE} - w) \xrightarrow{d} N(0, I(w)^{-1}), \tag{142}$$

*where $I(w)^{-1}$ is the Fisher information matrix at $w$ and $\xrightarrow{d}$ represents convergence in distribution.*

**Definition 51** (Minimax Decision Rule). *A decision rule $\hat{w}'$ is said to be minimax if it minimize the maximum expected cost, meaning*

$$\hat{w}' \equiv \inf_{\hat{w}} \sup_{w \in \Omega_W} \mathbb{E}[C(\hat{w}, w)|w, D]. \tag{143}$$

**Theorem 19** (Mean Squared Error (MSE)). *The expectation of the quadratic cost function (definition (53)) can be written*

$$\begin{aligned} \mathbb{E}[C(\hat{w}, w)|w, D] &= \mathbb{E}[(\hat{w} - w)^2|w, D] \\ &= \mathbb{E}[(\hat{w} - \mathbb{E}[\hat{w}])^2] + (w - \mathbb{E}[\hat{w}])^2 \\ &= Var[\hat{w}] + Bias[\hat{w}]^2 \end{aligned} \tag{144}$$

*where conditions have been suppressed in the second line (to fit to the page) and the bias of the estimator of $\hat{w}$ is defined viz*

$$Bias[\hat{w}] \equiv w - \mathbb{E}[\hat{w}|w, D]. \tag{145}$$

*If* $\mathbb{E}[C(\hat{w}, w)|w, D] \xrightarrow{data\to\infty} 0$ *then* $C(\hat{w}, w)$ *is a weakly consistent estimate of* $w$*. There can be different consistent estimates that converge towards* $w$ *at different speeds. It is desirable for an estimate to be consistent and with small (quadratic) cost, meaning that both the bias and variance of the estimator should be small. In many cases, however, there is bias-variance which means that both cannot be minimized at the same time.*

**Corollary 2** (MLE is Approximately Minimax for quadratic Loss)**.** *Under certain regularity conditions, the Maximum Likelihood decision rule (MLE)* $\hat{w}_{\text{MLE}}$ *is approximately minimax for the quadratic cost function (definition 53), meaning it approximately minimizes the maximum expected cost.*

*Proof.* From theorem 19

$$\mathbb{E}[(\hat{w} - w)^2] = \text{Var}[\hat{w}] + \text{Bias}[\hat{w}]^2. \tag{146}$$

Under the regularity conditions where the MLE is unbiased and has asymptotically minimal variance, the bias term vanish, meaning $\text{Bias}[\hat{w}_{\text{MLE}}] = 0$ and the variance term $\text{Var}[\hat{w}_{\text{MLE}}]$ is minimized among a class of estimators. Thus, the expected quadratic cost for the MLE can be approximated by

$$\begin{aligned}\mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2] &\approx \text{Var}[\hat{w}_{\text{MLE}}] \\ &\approx \frac{\text{tr}[I(w)^{-1}]}{n},\end{aligned} \tag{147}$$

where theorem 18 was used for the second line. The Cramer-Rao lower bound [27] for variance states that

$$\text{Var}[\hat{w}] \geq \frac{\text{tr}[I(w)^{-1}]}{n}, \tag{148}$$

implying that the MLE decision rule acheives the smallest possible variance asymptotically and therefore that

$$\sup_{w \in \Omega_W} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2] \approx \inf_{\hat{w}} \sup_{w \in \Omega_W} \mathbb{E}[(\hat{w} - w)^2], \qquad (149)$$

meaning the MLE decision rule is approximately the minimax decision rule under quadractic cost. □

**Example 8.1.**

*The bias-variance decomposition (theorem 19) is only relevant for frequentist statistics and Bayesian statistics does not struggle with this tradeoff. It relates to overfitting and underfitting. Bayesians do not fit in the same way as frequentists do. They do not determine a single set of parameters, rather they use a set of parameters due to integration. In that sense, they are protected against overfitting and underfitting and they do not struggle with hyperparameter finetuning as well.*

**Example 8.2.**

Consider $X_1, \dots, X_n \sim Ber(w)$. Determine the (quadratic) cost of three different decision rules for the mean; the arithmetic sample mean, the number 0.5 and the first data entry and $X_1$.

- *For the arithmetic mean*

$$\hat{w} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (150)$$

*meaning*

$$\begin{aligned} \mathbb{E}[\hat{w}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= w, \\ Var[\hat{w}] &= \frac{1}{n^2}\sum_{i=1}^{n}Var[X_i] \\ &= \frac{w(1-w)}{n}, \\ \mathbb{E}[(\hat{w}-w)^2] &= \frac{w(1-w)}{n}. \end{aligned} \tag{151}$$

- *For the number* 0.5

$$\hat{w} = 0.5 \tag{152}$$

*meaning*

$$\begin{aligned} \mathbb{E}[\hat{w}] &= 0.5, \\ Var[\hat{w}] &= 0, \\ \mathbb{E}[(\hat{w}-w)^2] &= (0.5-w)^2. \end{aligned} \tag{153}$$

- *For the first entry,* $X_1$*,*

$$\hat{w} = \frac{1}{n}\sum_{i=1}^{n}X_i \tag{154}$$

*meaning*

$$\begin{aligned} \mathbb{E}[\hat{w}] &= \mathbb{E}[X_1] \\ &= w, \\ Var[\hat{w}] &= Var[X_1] \\ &= w(1-w), \\ \mathbb{E}[(\hat{w}-w)^2] &= w(1-w). \end{aligned} \tag{155}$$

*The arithmetic mean minimizes the quadratic cost over the entire range of $w$, while the constant value $0.5$ performs better for a specific range of $w$. The cost for $X_1$ is independent of $n$, making it less favorable as $n$ increases.*

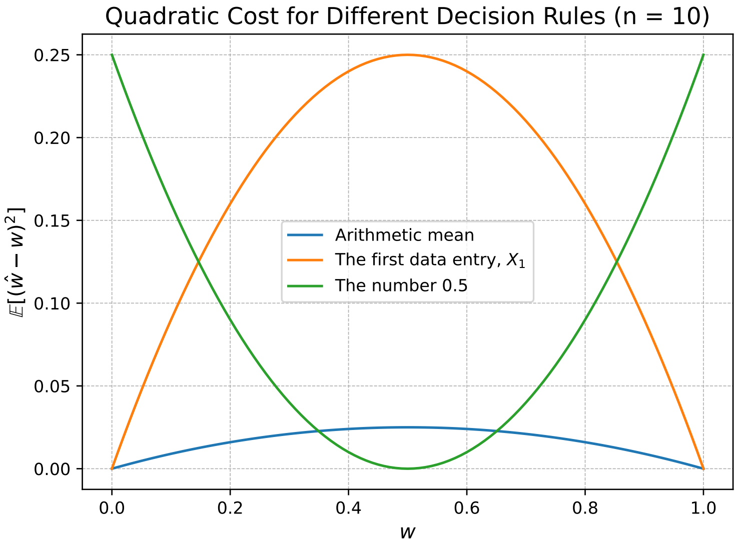

Figure 9: The quadratic cost, $\mathbb{E}[(\hat{w} - w)^2]$, for three different decision rules: the arithmetic mean (blue), the first data entry $X_1$ (orange), and the constant value 0.5 (green).

**Example 8.3.**

Determine the maximum likelihood estimate of $w$ for the model $(\{0,1\}, \{Ber(w)\}_{w \in (0,1)})$.

*In this case*

$$p(D_s | D_x, w) = \prod_{i=1}^{n} w^{x_i} (1-w)^{1-x_i}. \tag{156}$$

*Let $l(w) \equiv lnp(D_s|D_x, w)$, then*

$$\begin{aligned}\underset{w}{\operatorname{argmax}}\, l(w) &= \underset{w}{\operatorname{argmax}}\, p(D_s|D_x, w) \\ &= \underset{w}{\operatorname{argmax}} \ln \left( \prod_{i=1}^{n} w^{x_i} (1-w)^{1-x_i} \right) \\ &= \underset{w}{\operatorname{argmax}} \left[ \ln w \sum_{i=1}^{n} x_i + \ln(1-w) \sum_{i=1}^{n} (1-x_i) \right]\end{aligned} \tag{157}$$

*Now*

$$\frac{d}{dw} l(w) = \frac{\sum_{i=1}^{n} x_i}{w} - \frac{n - \sum_{i=1}^{n} x_i}{1-w} \tag{158}$$

*Requiring the derivative to vanish means the maximum likelihood estimate of w is given by*

$$\hat{w}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{159}$$

**Example 8.4.**

Determine the maximum likelihood estimate of $w$ for the model $([0, \infty), \{\text{Exp}(w)\}_{w>0})$.

*In this case*

$$p(D_s|D_x, w) = \prod_{i=1}^{n} w e^{-w x_i}. \tag{160}$$

*Let $l(w) \equiv lnp(D_s|D_x, w)$, then*

$$\frac{d}{dw} l(w) = \frac{n}{w} - \sum_{i=1}^{n} x_i \tag{161}$$

*Requiring the derivative to vanish means the maximum likelihood estimate of w is given by*

$$\hat{w}_{MLE} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} x_i}. \tag{162}$$

# Part III

# BAYESIAN STATISTICS

# CHAPTER 9

# Bayesian Statistics Introduction

Bayesian statistics is based on definition 43, where probability is viewed as a subjective degree of belief in the likelihood of an event occurring and parameters are modeled as random variables. The Bayesian framework originally come from the work of Bayes [28] and Laplace [29] with much of the modern discussions and formalism created later by Finetti [30] and Jeffreys [31] and Savage [32]. Although Bayesian statistics adopts a subjectivistic interpretation of probability, given established information the derived probabilities follow deductively from probability theory.

Following section **??**, statistics is framed as a game between a Robot and Nature in which each make decisions and the Robot is penalized for deviations from Natures actions via a cost function. The goal of the Robot is to minimize the expected cost associated with its decisions. To achieve this goal, the Robot utilizes its observations $x, D$ and background information $I$ to inform its decision-making process. Treating Natures decision $s \in \Omega_S$ as a random variable, the conditional expected cost, given the observations $D = \{(X_i = x_i, S_i = s_i)|i : n\}$, can be written [2]

$$\begin{aligned}\mathbb{E}[C(U,S)|I] &= \int dDdxdsC(U(x,D),s)p(X=x,S=s,D|I) \\ &= \int d\tilde{D}dsC(U(\tilde{D}),s)p(S=s,\tilde{D}|I)\end{aligned} \tag{163}$$

where $\tilde{D} = \{D, X = x\}$ and the Robot aims to find the decision rule which minimizes equation (163), meaning

$$U^* = \arg\min_U \mathbb{E}[C(U,S)|I]. \tag{164}$$

From theorem 10

$$\mathbb{E}[C(U,S)|I] = \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]]. \tag{165}$$

Using equation (165) in equation (164)

$$\begin{aligned} U^* &= \arg\min_U \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]] \\ &= \arg\min_U \int dx p(\tilde{D}|I)\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]. \end{aligned} \tag{166}$$

Since $p(\tilde{D}|I)$ is a non-negative function, the minimizer of the integral is the same as the minimizer of the conditional expectation, meaning

$$\begin{aligned} U^*(x) &= \arg\min_{U(x)} \mathbb{E}_{S|\tilde{D}}[C(U(x),S)|\tilde{D},I] \\ &= \arg\min_{U(x)} \int ds C(U(x,D),s)p(S=s|X=x,D,I). \end{aligned} \tag{167}$$

The probability $p(S=s|X=x,D,I)$ depend on the parameters $w_1, \dots w_n$ of the statistical model. Introducing the shorthand notation $W = w_1 \dots W = w_n \to w$, $dw_1 \dots dw_n \to dw$ and $X = x \to x$, then

$$\begin{aligned} p(s|x,D,I) &= \int dw p(w,s|x,D,I) \\ &= \int dw p(s|w,x,D,I)p(w|x,D,I) \end{aligned} \tag{168}$$

**Example 9.1.**

*Writing out the shorthand notation*

$$\begin{aligned} p(W = w_1, \dots, W = w_n, S = s|X = x, D, I) &\to p(w,s|x,D,I), \\ dw_1 \dots dw_n &\to dw. \end{aligned}$$

(169)

---

To evaluate $p(w|D, I)$ a combination of marginalization, Bayes' theorem, and the chain rule (see chapter 3) can be employed viz

$$\begin{aligned} p(w|x, D, I) &= p(w|D, I) \\ &= \frac{p(D_s|w, D_x, I)p(w|I)}{p(D_s|D_x, I)}, \end{aligned} \tag{170}$$

where $D_s = \{S = s_1 \ldots S = s_n\}$, $D_x = \{X = x_1, \ldots X = x_n\}$ and $p(D_s|D_x, I)$ can be expanded via marginalization and axiom 6 has been used for the first and second equality.

**Axiom 6** (Relevance of Observations)**.** *The Robot's observations are relevant for estimating Nature's model only when they map to known actions of Nature.*

$p(w|I)$ is the Robot's prior belief about Nature's actions for $w$. $p(D_s|w, D_x, I)$ is the likelihood of the past observations of Nature's actions, and $p(w|D, I)$ called the posterior distribution represent the belief of the Robot after seeing data. The prior distribution depends on parameters that must be specified and cannot be learned from data since it reflects the Robot's belief before observing data. These parameters are included in the background information, $I$. From equation (170), it is evident that, given the relevant probability distributions are specified, the probability of a parameter taking a specific value follows deductively from probability theory. The subjectivity arises from the assignment and specification of probability distributions which depend on the background information.

**Example 9.2.**

*In general the random variable X represent the observations the Robot has available that are related to the decision Nature is going* *to make. However, this information may not be given, in which case $D_x = \emptyset$ and consequently $D = D_s$. In this case, the Robot is forced to model the decisions of Nature with a probability distribution with associated parameters without observations. From equation* (274) *the optimal action for the Robot can be written*

$$U^* = \arg\min_U \mathbb{E}[C(U,S)|D_s, I]. \tag{171}$$

---

# CHAPTER 10

# Assigning a Cost Function

The cost function (see definition 44) associates a numerical penalty to the Robot's action and thus the details of it determine the decisions made by the Robot. Under certain conditions, a cost function can be shown to exist [3], however, there is no systematic way of producing or deriving the cost function beyond applied logic. In general, the topic can be split into considering a continuous and discrete action space, $\mathbb{U}$.

## 10.1 CONTINUOUS ACTION SPACE

In case of a continuous action space, the cost function is typically picked from a set of standard choices.

**Definition 52** (Linear Cost Function)**.** *The linear cost function is defined viz*

$$C(U(x),s) \equiv |U(x)-s|. \tag{172}$$

**Theorem 20** (Median Estimator)**.** *The median estimator follows from assuming linear loss viz*

$$\begin{aligned}\mathbb{E}_{S|X}[C(U(X),S)|x,D,I] &= \int_{-\infty}^{\infty} ds|U(x)-s|p(s|x,D,I)\\ &= \int_{-\infty}^{U(x)}(s-U(x))p(s|x,D,I)ds\\ &\quad+\int_{U(x)}^{\infty}(U(x)-s)p(s|x,D,I)ds\end{aligned}$$

$$(173)$$

$$\begin{aligned}0 &= \frac{d\mathbb{E}_{S|X}[C(U(X),S)|x,D,I]}{dU(X)}\bigg|_{U(x)=U^*(x)} \\ &= (U^*(x)-U^*(x))p(U^*(x)|x,D,I) + \int_{-\infty}^{U^*(x)} p(s|x,D,I)ds \\ &\quad + (U^*(x)-U^*(x))p(U^*(x)|x,D,I) - \int_{U^*(x)}^{\infty} p(s|x,D,I)ds\end{aligned} \tag{174}$$

$$\begin{aligned}\int_{-\infty}^{U^*(x)} p(s|x,D,I)ds &= \int_{U^*(x)}^{\infty} p(s|x,D,I)ds \\ &= 1 - \int_{-\infty}^{U^*(x)} p(s|x,D,I)ds\end{aligned} \tag{175}$$

$$\int_{-\infty}^{U^*(x)} p(s|x,D,I)ds = \frac{1}{2} \tag{176}$$

*which is the definition of the median.*

**Definition 53** (Quadratic Cost Function)**.** *The quadratic cost function is defined as*

$$C(U(x),s) \equiv (U(x)-s)^2. \tag{177}$$

**Theorem 21** (Expectation value). *The expectation value follows from assuming quadratic loss viz*

$$\begin{aligned}\mathbb{E}_{S|X}[C(U(X),S)|x,D,I] &= \int ds(U(x)-s)^2p(s|x,D,I)\\ &\Downarrow\\ \frac{d\mathbb{E}_{S|X}[C(U(X),S)|x,D,I]}{dU(X)}\bigg|_{U(x)=U^*(x)} &= 2U^*(x)-2\int dssp(s|x,D,I)\\ &= 0\\ &\Downarrow\\ U^*(x) &= \int dssp(s|x,D,I)\\ &= \mathbb{E}[S|x,D,I]\end{aligned} \tag{178}$$

*which is the definition of the expectation value.*

**Definition 54** (0-1 Cost Function). *The 0-1 cost function is defined viz*

$$C(U(x),s) \equiv 1-\delta(U(x)-s). \tag{179}$$

**Theorem 22** (MAP). *The maximum aposteriori (MAP) follows from assuming 0-1 loss viz*

$$\begin{aligned}\mathbb{E}_{S|X}[C((X),S)|x,D,I] &= 1-\int ds\delta(U(x)-s)p(s|x,D,I)\\ &\Downarrow\\ \frac{d\mathbb{E}_{S|X}[C(U(X),S)|x,D,I]}{dU(X)}\bigg|_{U(x)=U^*(x)} &= -\frac{dp(s|x,D,I)}{ds}\bigg|_{s=U^*(x)}\\ &= 0\end{aligned} \tag{180}$$

*which is the definition of the MAP.*

### 10.1.1 *Regression*

Regression involves the Robot building a model, $f : \mathbb{W} \times X \mapsto \mathbb{R}$, with associated parameters $\theta \in \mathbb{W}$, that estimates Nature's actions $S$ based on observed data $X$. Note that the output of $f$ is $\mathbb{R}$ implying that $S$ is assumed continuos. The model $f$ acts as a proxy for the Robot in that it on behalf of the Robot estimates the action of Nature given an input. Hence, in providing an estimate, the model must make a choice, similar to the Robot and thus the Robot must pick a cost function for the model. In this study, the quadratic cost function 53 will be considered to review the subject. From theorem 21 the best action for the Robot can be written

$$U^*(x) = \int dssp(s|x,D,I) \tag{181}$$

Assuming the actions of Nature follow a normal distribution with the function $f$ as mean and an unknown variance, $\xi \in \mathbb{W}$

$$p(s|x,\theta,\xi,I) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(\theta,x)-s)^2}. \tag{182}$$

Using equation (182) and marginalizing over $\xi, \theta$

$$\begin{aligned} p(s|x,D,I) &= \int p(s,\theta,\xi|x,D,I)d\theta d\xi \\ &= \int p(s|x,\theta,\xi,D,I)p(\theta,\xi|x,D,I)d\theta d\xi \quad (183) \\ &= \int p(s|x,\theta,\xi,I)p(\theta,\xi|D,I)d\theta d\xi, \end{aligned}$$

where it has been used that $p(s|\theta,\xi,x,D,I) = p(s|\theta,\xi,x,I)$ since by definition $f$ produce a $1-1$ map of the input $x$

(equation (182)) and $p(\theta,\xi|x,D,I) = p(\theta,\xi|D,I)$ from axiom 6. Using equation (183) in equation (181)[1]

$$\begin{aligned} U^*(x) &= \int f(\theta,x)p(\theta,\xi|D,I)d\theta d\xi, \\ &= \mathbb{E}[f|x,D,I] \end{aligned} \tag{184}$$

where it has been used that

$$\begin{aligned} \mathbb{E}[S|x,\theta,\xi,I] &= \int sp(s|x,\theta,\xi,I)dy \\ &= f(\theta,x) \end{aligned} \tag{185}$$

according to equation (182). Using Bayes theorem

$$p(\theta,\xi|D,I) = \frac{p(D_s|D_x,\theta,\xi,I)p(\theta,\xi|D_x,I)}{p(D_s|D_x,I)} \tag{186}$$

where from marginalization

$$p(D_s|D_x,I) = \int p(D_s|D_x,\theta,\xi,I)p(\theta,\xi|D_x,I)d\theta d\xi. \tag{187}$$

Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written (using equation (182))

$$p(D_s|D_x,\theta,\xi,I) = \left(\frac{\xi}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{\xi}{2}(f(\theta,x_i)-s_i)^2} \tag{188}$$

From the chain rule (see theorem 1) and axiom 6

$$p(\theta,\xi|D_x,I) = p(\theta|\xi,I)p(\xi|I). \tag{189}$$

Assuming the distributions over $W_\theta$ are i) independent of $\xi$ and ii) normally distributed[2] with zero mean and a precision described by a hyperparameter, $\lambda$.

$$\begin{aligned} p(\theta|\xi,I) &= p(\theta|I) \\ &= \int p(\theta|\lambda,I)p(\lambda|I)d\lambda \end{aligned} \tag{190}$$

[1] Note that a function of a random variable is itself a random variable, so $f$ is a random variable.

[2] The normally distributed prior is closely related to weight decay [33], a principle conventionally used in frequentist statistics to avoid the issue of overfitting.

The precision is constructed as a wide gamma distribution so as to approximate an objective prior

$$p(\theta|\lambda, I)p(\lambda|I) = \prod_{q=1}^{\tilde{n}} \frac{\lambda_q^{\frac{n_q}{2}}}{(2\pi)^{\frac{n_q}{2}}} e^{-\frac{\lambda_q}{2}\sum_{l=1}^{n_q}\theta_l^2} \frac{\beta_q^{\alpha_q}}{\Gamma(\alpha_q)} \lambda_q^{\alpha_q - 1} e^{-\beta_q \lambda_q} \tag{191}$$

where $\alpha_q, \beta_q$ are prior parameters (a part of the background information) and $\tilde{n}$ is the number of hyper parameters. In the completely general case $\tilde{n}$ would equal the number of parameters $\theta$, such that each parameter has an independent precision. In practice, the Robot may consider assigning some parameters the same precision, e.g. for parameters in the same layer in a neural network. Since $p(\xi|I)$ is analogous to $p(\lambda|I)$ – in that both are prior distributions for precision parameters – $p(\xi|I)$ is assumed to be a wide gamma distribution, then

$$\begin{aligned} p(\xi|I) &= \text{Ga}(\xi|\tilde{\alpha}, \tilde{\beta}) \\ &= \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \xi^{\tilde{\alpha}-1} e^{-\tilde{\beta}\xi}. \end{aligned} \tag{192}$$

At this point equation (181) is fully specified (the parameters $\alpha, \beta, \tilde{\alpha}, \tilde{\beta}$ and the functional form of $f(\theta, x)$ are assumed specified as part of the background information) and can be approximated by obtaining samples from $p(\theta, \xi, \lambda|D, I)$ via HMC [34–37] (see appendix A for a review of HMC). The centerpiece in the HMC algorithm is the Hamiltonian defined viz [36, 37]

$$H \equiv \sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \ln[p(\theta, \xi, \lambda|D, I)] + const, \tag{193}$$

where

$$p(\theta, \xi|D, I) = \int d\lambda p(\theta, \xi, \lambda|D, I). \tag{194}$$

Besides its function in the HMC algorithm, the Hamiltonian represent the details of the Bayesian model well and should be a familiar sight for people used to the more commonly applied frequentist formalism (since, in this case, it is in form similar to a cost function comprised of a sum of squared errors, weight decay on the coefficients and further penalty terms [38–40]). Using equations (186)-(194) yields

$$\begin{aligned} H = {} & \sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} + \frac{n}{2}[\ln(2\pi) - \ln(\xi)] + \frac{\xi}{2} \sum_{i=1}^{n} (f(\theta, x_i) - s_i)^2 \\ & + \sum_{q=1}^{\tilde{n}} \Bigg( \ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln(\lambda_q) + \beta_q \lambda_q \\ & \qquad + \frac{n_q}{2} (\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} \theta_l^2 \Bigg) \\ & + \ln(\Gamma(\tilde{\alpha})) - \tilde{\alpha} \ln(\tilde{\beta}) + (1 - \tilde{\alpha}) \ln(\xi) + \tilde{\beta} \xi + const. \end{aligned} \tag{195}$$

## 10.2 DISCRETE ACTION SPACE

In case of a continuous action space, the conditional expected loss can be written

$$\mathbb{E}_{S|X}[C(U(X), S)|x, D, I] = \sum_{s \in \mathbb{S}}^{n} C(U(x), s) p(s|x, D, I), \tag{196}$$

where the cost function is typically represented in matrix form viz

$$\begin{array}{cc|ccc|}
\multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{3}{c}{S} \\
\multicolumn{1}{c}{} & \multicolumn{1}{c}{} & s_1 & \dots & \multicolumn{1}{c}{s_{\dim(\mathbb{S})}} \\ \cline{3-5}
U(x) & u_1 & C(u_1, s_1) & \dots & C(u_1, s_{\dim(\mathbb{S})}) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & u_{\dim(\mathbb{U})} & C(u_{\dim(\mathbb{U})}, s_1) & \dots & C(u_{\dim(\mathbb{U})}, s_{\dim(\mathbb{S})}) \\ \cline{3-5}
\end{array}$$

### 10.2.1 *Classification*

Classification is the discrete version of regression, meaning it involves the Robot building a model, $f : \mathbb{W} \times X \mapsto [0,1]$, with associated parameters $\theta \in \mathbb{W}$, that estimates Nature's actions $S$ based on observed data $X$. As opposed to regression, the random variable $S$ is now discrete and the function is identified with the probability of each action

$$p(S = s|x, \theta, I) = f_{S=s}(\theta, x), \tag{197}$$

with

$$\sum_{s \in \mathbb{S}} p(S = s|x, \theta, I) = 1. \tag{198}$$

In this case, the Robot's action space is equal to Natures action space, with the possible addition of a reject option, $\mathbb{U} = \mathbb{S} \cup \text{"Reject"}$. To reivew this subject the Robot will be considered to be penalized equally in case of a classification error, which corresponds to the $0-1$ cost function, with the addition of a reject option at cost $\lambda$. This means

$$C(U(x), s) = 1 - \delta_{U(x),s} + (\lambda - 1)\delta_{U(x),\text{"Reject"}}. \tag{199}$$

The optimal decision rule for the robot can the be written

$$\begin{aligned} U^*(x) &= \arg\min_{U(x)} \mathbb{E}[C(U(X), S)|x, D, I] \\ &= \arg\min_{U(x)} \bigg( \sum_s C(U(x), s) p(S = s|x, D, I) \\ &\qquad\qquad + (\lambda - 1)\delta_{U(x),\text{"Reject"}} \bigg) \\ &= \arg\min_{U(x)} \bigg( 1 - p(S = U(x)|x, D, I) \\ &\qquad\qquad + (\lambda - 1)\delta_{U(x),\text{"Reject"}} \bigg). \end{aligned} \tag{200}$$

In absence of the reject option, the optimal decision rule is to pick the MAP, similar to theorem 22. Using equation (276) and marginalizing over $\theta$

$$\begin{aligned} p(S = U(x)|x, D, I) &= \int p(S = U(x), \theta|x, D, I)d\theta \\ &= \int p(S = U(x)|x, \theta, D, I)p(\theta|x, D, I)d\theta \\ &= \int p(S = U(x)|x, \theta, I)p(\theta|D, I)d\theta \\ &= \int f_{S=U(x)}(\theta, x)p(\theta|D, I)d\theta \\ &= \mathbb{E}[f_{S=U(x)}(\theta, x)|D, I], \end{aligned} \tag{201}$$

where for the second to last equality it has been assumed that $p(S = U(x)|\theta, x, D, I) = p(S = U(x)|\theta, x, I)$ since by definition $f$ (see equation (276)) produce a $1-1$ map of the input $x$ and $p(\theta|x, D, I) = p(\theta|D, I)$ from axiom 6. From Bayes theorem

$$p(\theta|D, I) = \frac{p(D_s|D_x, \theta, I)p(\theta|D_x, I)}{p(D_s|D_x, I)}, \tag{202}$$

where from axiom 6 $p(\theta|D_x, I) = p(\theta|I)$. Assuming the distribution over $\theta$ is normally distributed with zero mean and a precision described by a hyperparameter, $\lambda$,

$$p(\theta|I) = \int p(\theta|\lambda, I)p(\lambda|I)d\lambda. \tag{203}$$

where $p(\theta|\lambda, I)p(\lambda|I)$ is given by equation (281). Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written [41]

$$\begin{aligned} p(D_s|D_x, \theta, I) &= \prod_{i=1}^{n} p(S = s_i|X = x_i, \theta, I) \\ &= \prod_{i=1}^{n} f_{s_i}(\theta, x_i) \end{aligned} . \tag{204}$$

At this point equation (200) is fully specified and can be approximated by HMC similarly to the regression case. In this case, the model can be represented by the Hamiltonian

$$H \equiv \sum_q \sum_l \frac{p_l^2}{2m_l} - \ln(p(\theta, \lambda|D, I)) + const \tag{205}$$

where

$$p(\theta|D, I) = \int d\lambda p(\theta, \lambda|D, I). \tag{206}$$

Using equations (275)-(282) in equation (283) yields the Hamiltonian

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \sum_{i=1}^{n} \ln(f_{s_i}(\theta, x_i)) + \text{const} \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg( \ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1-\alpha_q)\ln(\lambda_q) + \beta_q \lambda_q. \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} \theta_l^2 \Bigg) \end{aligned} \tag{207}$$

Sampling equation (285) yields a set of coefficients which can be used to compute $\mathbb{E}[f_s(\theta, x)|D, I]$ which in turn (see euation (275)) can be used to compute $U^*(x)$.

**Example 10.1.**

*Let $\xi \equiv e^{\zeta}$, such that $\zeta \in [-\infty, \infty]$ maps to $\xi \in [0, \infty]$ and $\xi$ is ensured to be positive definite regardless of the value of $\zeta$. Using the differential $d\xi = \xi d\zeta$ in equation (184) means $p(\theta, \xi, \lambda|D, I)$ is multiplied with $\xi$. Hence, when taking $-\ln(p(\theta, \xi, \lambda|D, I))$ according to equation (193), a $-\ln(\xi)$ is added to the Hamiltonian. In practice this means*

$$(1-\tilde{\alpha})\ln(\xi) \in H \Rightarrow -\tilde{\alpha}\ln(\xi). \tag{208}$$

---

**Example 10.2.**

$$C(U(x), s) = \alpha \cdot swish(U(x) - s, \beta) + (1-\alpha) \cdot swish(s - U(x), \beta) \tag{209}$$

*where*

$$swish(z, \beta) = \frac{z}{1 + e^{-z\beta}}. \tag{210}$$

*and $z \equiv U(x) - s$. Taking $\alpha \ll 1$, then $z < 0$ will be penalized relatively more than $z > 0$. $z < 0$ corresponds to underestimation, so this is penalized greater relative to overestimation. Now*

$$\mathbb{E}[C|\dots] = \int dsp(s|\dots)\Bigg(\alpha \cdot swish(U(x) - s, \beta) + (1-\alpha) \cdot swish(s - U(x), \tag{211}$$

*Let $z \equiv U(x) - s$, then*

$$\begin{aligned}
\frac{dC}{dU(x)} &= \frac{dC}{dz}\frac{dz}{dU(x)} \\
&= \left(\frac{\alpha}{1+e^{-\beta z}} - \frac{1-\alpha}{1+e^{\beta z}} + \frac{\alpha\beta e^{-\beta z} z}{(1+e^{-\beta z})^2} + \frac{(1-\alpha)\beta e^{\beta z} z}{(1+e^{\beta z})^2}\right)\frac{dz}{dU(x)} \\
&= \frac{\beta z e^{\beta z} - e^{\beta z} - 1}{(1+e^{\beta z})^2} + \alpha + \mathcal{O}(\alpha^2) \\
&\approx \alpha - \frac{1}{(1+e^{\beta z})^2}
\end{aligned}$$

(212)

$$
\begin{aligned}
\frac{d\mathbb{E}[C|\ldots]}{dU(x)} &\approx \int dsp(s|\ldots)\left(\alpha - \frac{1}{(1+e^{\beta z})^2}\right) \\
&= \alpha - \int dsp(s|\ldots)\frac{1}{(1+e^{\beta z})^2} \\
&= 0
\end{aligned}
\tag{213}
$$

*$\frac{1}{(1+e^{\beta z})^2}$ approximate a unit step which is $1$ for $z < 0$ and $0$ otherwise. $z < 0 \Rightarrow s > U(x)$. This means*

$$\int_{-\infty}^{\infty} dsp(s|\ldots)\frac{1}{(1+e^{\beta z})^2} \approx \int_{U(x)}^{\infty} dsp(s|\ldots) \tag{214}$$

*This means*

$$\alpha \approx \int_{U(x)}^{\infty} dsp(s|\ldots) \tag{215}$$

---

**Example 10.3.**

*Suppose there is a game between a Robot and Nature in which the Robot objective is to guess the position of the Robot. The Robot is given measurements of its velocity and acceleration, at any time step and must formulate a belief about its position. Nature decides the true position of the Robot and will penalize the Robot according to the deviation between the Robots estimate of its position and the true position viz*

$$C(U(x), s) = (U(x) - s)^2, \tag{216}$$

*where $s$ is the true position $U(x)$ is the Robots estimate based on data $x$ containing velocity and acceleration measurements.*

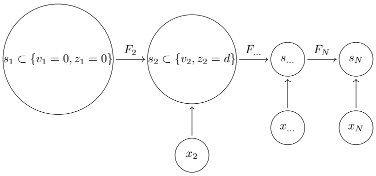
$s_1 \subset \{v_1 = 0, z_1 = 0\}$
$F_2$
$s_2 \subset \{v_2, z_2 = d\}$
$F_{\dots}$
$s_{\dots}$
$F_N$
$s_N$
$x_2$
$x_{\dots}$
$x_N$

Figure 10

*Suppose the Robot is given all historical data, then the expected cost can be written*

$$\mathbb{E}_{S|X}[C(U(X), S|s_1, \{x\}_{2:N}, I)] = \int ds (U(x_N) - s)^2 p(s|s_1, \{x\}_{2:N}, I). \tag{217}$$

*The optimal decision is defined by*

$$\frac{d}{dU(X)}\left(\mathbb{E}_{S|X}[C(U(X), S|s_1, \{x\}_{2:N}, I)]\right)\Bigg|_{U(x_N)=U^*(x_N)} = 0 \tag{218}$$

*leading to*

$$U^*(x_N) = \mathbb{E}[S_N|s_1, \{x\}_{2:N}, I]. \tag{219}$$

*The inutitive interpretation of equation* (219) *is that the optimal decision of the Robot is to estimate the position that it expects Nature to have chosen. Now*

$$\mathbb{E}[S_N|s_1, \{x\}_{2:N}, I] = \int ds_N s_N p(s_N|s_1, \{x\}_{2:N}, I). \tag{220}$$

*Assume that*

$$\begin{aligned} p(s_i|s_{i-1}, I) &= N(s_i|\mu_i = F_i s_{i-1} + b_i, \Sigma_i = Q_i), \\ p(x_i|s_i, I) &= N(x_i|\mu_i = H_i s_i + d_i, \Sigma_i = R_i) \end{aligned} \tag{221}$$

*Hence,* $p(s_N|s_1, \{x\}_{2:N}, I)$ *must be reformulated such that the above can be utilized. Now*

$$\begin{aligned} p(s_N|s_1, \{x\}_{2:N}, I) &= \frac{p(x_N|s_N, s_1, \{x\}_{2:N-1}, I)p(s_N|s_1, \{x\}_{2:N-1}, I)}{p(x_N|s_1, \{x\}_{2:N-1}, I)} \\ &= \frac{p(x_N|s_N, I)p(s_N|s_1, \{x\}_{2:N-1}, I)}{p(x_N|s_1, \{x\}_{2:N-1}, I)} \end{aligned}$$

$$\tag{222}$$

*where*

$$\begin{aligned} p(s_N|s_1,\{x\}_{2:N-1},I) &= \int ds_{N-1} p(s_N, s_{N-1}|s_1,\{x\}_{2:N-1},I) \\ &= \int ds_{N-1} p(s_N|s_{N-1},s_1,\{x\}_{2:N},I) p(s_{N-1}|s_1,\{x\}_{2:N-1}, \\ &= \int ds_{N-1} p(s_N|s_{N-1}) p(s_{N-1}|s_1,\{x\}_{2:N-1},I) \\ &= \int ds_{N-1} N(s_N|F_N s_{N-1}+b_N, Q_N) N(s_{N-1}|\mu_{N-1}, \Sigma_{N-} \\ &= N(s_N|\mu_{N|N-1}, \Sigma_{N|N-1}) \end{aligned} \tag{223}$$

*where*

$$\begin{aligned} \mu_{N|N-1} &= F_N \mu_{N-1} + b_N \\ \Sigma_{N|N-1} &= F_N \Sigma_{N-1} F_N^T + Q_N \end{aligned} \tag{224}$$

*Using equation* (223) *and* (221) *in equation* (222) *then yields [2]*

$$p(s_N|s_1,\{x\}_{2:N},I) = N(s_N|\mu_N,\Sigma_N) \tag{225}$$

*where*

$$\begin{aligned} \mu_N &= \mu_{N|N-1} + K_N(x_N - \hat{x}_N), \\ \Sigma_N &= \Sigma_{N|N-1} - K_N S_N K_N^T, \\ K_N &= \Sigma_{N|N-1} H_N^T S_N^{-1}, \\ \hat{x}_N &= H_N \mu_{N|N-1} + d_N, \\ S_N &= H_N \Sigma_{N|N-1} H_N^T + R_N \end{aligned} \tag{226}$$

*Combining equation* (225) *with* (220) *then yields the optimal decision for the Robot*

$$\mathbb{E}[S_N|s_1,\{x\}_{2:N},I] = \mu_N. \tag{227}$$

**Example 10.4.**

*Consider a Robot moving in one dimension. Every time interval $\Delta t$, the Robot will sample a wheel counter and an accelerometer. The wheel counter will be incremented every distance $d$. The Robot is interested in knowing its position and velocity. Expand the position viz*

$$z(t+\Delta t) = z(t) + \Delta t \frac{dz(t)}{dt} + \frac{1}{2}(\Delta t)^2 \frac{d^2 z(t)}{dt^2} + \mathcal{O}(\Delta t^3) \quad (228)$$

*which discretize to*

$$z_k \simeq z_{k-1} + \Delta t v_{k-1} + \frac{1}{2}(\Delta t)^2 a_{k-1}, \quad (229)$$

*where $\Delta t = const$ and*

$$\begin{aligned} v_k &\simeq v_{k-1} + \Delta t a_{k-1}, \\ a_k &\simeq a_{k-1}, \end{aligned} \quad (230)$$

*This means*

$$\begin{aligned} s_k &= \begin{pmatrix} z_k \\ v_k \\ a_k \end{pmatrix} \\ &\simeq \underbrace{\begin{pmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{pmatrix}}_{=F_k} \begin{pmatrix} z_{k-1} \\ v_{k-1} \\ a_{k-1} \end{pmatrix} \end{aligned} \quad (231)$$

*where $b_k = \emptyset$ for simplicity. Now take*

$$\begin{aligned} x_k &= \begin{pmatrix} c_k \\ a_k \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} d^{-1} & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{=H_k} \begin{pmatrix} z_{k-1} \\ v_{k-1} \\ a_{k-1} \end{pmatrix} + r_k \end{aligned} \quad (232)$$

*where $r_k \sim N(0, R_k)$ with*

$$R_k = \begin{pmatrix} \sigma_z^2 & \sigma_z\sigma_a \\ \sigma_z\sigma_a & \sigma_a^2 \end{pmatrix}, \tag{233}$$

*where $\sigma_z$ and $\sigma_a$ are estimated from observations. The process noise can be determined viz Kalman link*

$$Q_k = \begin{pmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} & \frac{\Delta t^2}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 & \Delta t \\ \frac{\Delta t^2}{2} & \Delta t & 1 \end{pmatrix} \sigma^2, \tag{234}$$

*where $\sigma \sim \Delta a$.*

**Example 10.5.**

*Specifically, a first attempt at quantifying what is called* Renzo's rule *[42] by using Bayesian artificial neural networks (BANNs) to detect anomalies in the baryonic and observed galactic rotation curves will be presented. Renzo's rule state that*

"every time there is a feature in the radial light distribution the rotation curve shows a corresponding feature"

*It has been corroborated by examining the quality of fits from scaling the contributions of baryonic matter in rotation curves [43–45] and is sometimes used as an argument in favor of modified gravitational dynamics (see e.g. [43, 45]).*
*The basic idea for quantifying Renzo's rule is to let two separate BANNs identify anomalies in the baryonic and observed galactic rotation curves such that the coincidence of the anomalies can be compared and analyzed. In theory, if Renzo's rule is completely accurate and the anomalies detected by the BANNs correspond to features in the baryonic and observed rotation curves, one would expect a one-to-one correspondence between the anomalies in the baryonic and observed rotation curves. Due to measurement imperfections in the independent measurements techniques of the*

*baryonic and observed rotation curves, it is, however, not expected that there be a complete one-to-one correspondence even if Renzo's rule is completely accurate and the anomalies detected by the BANNs correspond to features in the baryonic and observed rotation curves. However the method presented and the initial results motivate further investigation.*

## 10.3 DATA

*Data from the SPARC database with selection criteria identical to Paper III [46] is used in Paper VI [***Petersen:2020renzo***]. This means rotation curve data from* 152 *galaxies making up* 3143 *rotation curve points.*

## 10.4 CONSIDERATIONS REGARDING APPROACH

*In order to quantify Renzo's rule, the degree of coincidence of anomalies in the baryonic rotation curve,* $v_{bar}(r)$*, and the observed rotation curve,* $v_{obs}(r)$*, will be investigated in this study. In order to quantify anomalies as objectively and unbiased as possible, two autoencoder (AE) BANNs will be used – One (AE*$_{obs}$*) with input* $\{r, v_{obs}\}$ *and the other (AE*$_{bar}$*) with input* $\{r, v_{bar}\}$*. The autoencoder will construct a representation of the input and transform the input back and forth between the representation. After successful training on a set of training data, the autoencoder will be able to reconstruct test input that is similar to that of the training input. In order to test Renzo's rule as accurately as possible, it is desired that the autoencoder is trained on featureless data such that it will raise a metaphorical flag every time it encounters a feature in the test data. In order to avoid the bias associated with cherry picking featureless data for training, a set of* 152 *mock galaxies are generated using randomly perturbed parameters from the SPARC galaxies (to ensure the parameters used are different but realistic). This is done by sampling an exponential disk pro-*

*file [47] with noise corresponding to a* 5% *relative uncertainty*

$$v_{bar}^{(mock)} \sim \mathcal{N}(v_N, 0.05 v_N), \tag{235}$$

*where in equation* (235) *"$\sim$" denote "distributed as" and*

$$v_N = \sqrt{\frac{G_N m_d r^2}{2 r_d^3}\left(I_0\left(\frac{r}{2r_d}\right)K_0\left(\frac{r}{2r_d}\right) - I_1\left(\frac{r}{2r_d}\right)K_1\left(\frac{r}{2r_d}\right)\right)} \tag{236}$$

*with $G_N$ being Newtons gravitational constant, $m_d = 2\pi r_d^2 \Upsilon^d S_d$ the central surface density with the disc mass to light ratio $\Upsilon^d = 0.5\frac{m_\odot}{L_\odot}$ [48]. $S_d$ and $r_d$ is determined by randomly perturbing the corresponding measured quantities of the SPARC database. The exponential disk profile is sampled in the range $\{0, 10\} r_d$ with* 50 *data points. The mock data for $v_{obs}$ is generated using the Radial Acceleration Relation (RAR) [49] such that*

$$v_{obs}^{(mock)} \sim \mathcal{N}(v_{tot}, 0.05 v_{tot}), \tag{237}$$

*where in equation* (237) *"$\sim$" denote "distributed as" and*

$$v_{tot} = v_N \left(1 - e^{-\sqrt{\frac{v_N^2}{a_0 r}}}\right)^{-\frac{1}{2}} \tag{238}$$

*with $a_0 = 1.2 \cdot 10^{-10} \frac{m}{s^2}$. Using the training data from the* 152 *mock galaxies, a feature is defined as* the input the AE – trained on featureless mock galaxies – is not able to reconstruct accurately. *What constitutes "accurately" will be elaborated later on.*

*To keep things simple, both AEs will be on the form of two-layer perceptrons [38, 39] with sigmoid activation functions, meaning*

$$\begin{aligned} f_k^{(n)} &= b_k^{(3)} + \sum_{l=1}^{n_2} W_{kl}^{(3)} h_l^{(n,2)}, \\ h_l^{(n,2)} &= g\left( b_l^{(2)} + \sum_{m=1}^{n_1} W_{lm}^{(2)} h_m^{(n,1)} \right), \\ h_m^{(n,1)} &= g\left( b_m^{(1)} + \sum_{s=1}^{P} W_{ms}^{(1)} x_s^{(n)} \right), \end{aligned} \tag{239}$$

*where $b, W$ are coefficients (collectively denoted by $\theta$), $K = P$ for an AE and*

$$\begin{aligned} g(z) &= sig(z) \\ &= \frac{1}{1+e^{-z}}. \end{aligned} \tag{240}$$

*This choice of $g$ shrink the influence of lower level coefficients since*

$$\begin{aligned} \frac{dsig(z_j)}{dz_u} &= sig(z_j)(1-sig(z_j))\delta_{ju} \\ &\lesssim \frac{1}{4} \end{aligned} \tag{241}$$

*will be contained in the gradients used for training. This is the so-called "vanishing gradients problem" [50, 51]. By training the AE via Hamiltonian Monte Carlo (HMC) [35–37], the problem can be alleviated to some degree for small $z$ by adjusting the parameters of the algorithm (the masses). In order for $z$ to be small, the input must be normalized. Due to the lack of normality of the input data (the rotation curve data), data is normalized according to*

$$\begin{aligned} \{\check{r}, \check{v}_{bar}\} &\equiv \left\{ \frac{r - \min(r)}{\max(r) - \min(r)}, \frac{v_{bar} - \min(v_{bar})}{\max(v_{bar}) - \min(v_{bar})} \right\}, \\ \{\check{r}, \check{v}_{obs}\} &\equiv \left\{ \frac{r - \min(r)}{\max(r) - \min(r)}, \frac{v_{obs} - \min(v_{obs})}{\max(v_{obs}) - \min(v_{obs})} \right\}, \end{aligned}$$

(242)

*where the* max *and* min *operations go over all data points in the training/test data, respectively. The normalized radii and velocities for the test and training data are shown in figure* **??** *in appendix* **??***. The* min *and* max *operations are non-differentiable, which means the systematic uncertainties included in the variables (inclination, $\alpha$, distance, D, and the unitless mass to light ratios $\tilde{\Upsilon}$) cannot be marginalized over – as would otherwise be an obvious choice in the Bayesian formalism. Instead, the uncertainties of the input data are treated as hyper parameters, $\xi$, controlled by a set of vague priors in order to be as unbiased as possible. Due to the treatment of the uncertainties the data from different galaxies can be stacked "head to tail" in the input array, $x_k^{(n)}$, with k referring to the input variables and n the data entry. The target values, $y_k^{(n)}$, are equal to the input ($y_k^{(n)} = x_k^{(n)}$) for the AE, but to avoid confusion the notational distinction between the two are kept. In order for the AE to be able to detect structure in the rotation curves, the AE take* 5 *consecutive rotation curve points as input, meaning $k = 1, 2, \dots 10$ (since the input consists of* 5 *radial points and* 5 *velocity points) and $n = 1, 2, \dots 3138$ (note the endpoint is* $3143 - 5$*) for the test data.*

## 10.5 METHOD

*The idea is to feed $v_N$ to a two-layer perceptron and fit it to $v_N$ (an auto encoder). Then after this, the auto encoder will be fed $v_{bar}$ and it when behavior that deviates from the theoretical one is encountered, the model will flag this. This is done for both the baryonic and observed curved. Then the positioning of anomalies can be compared.*

*Perhaps we can do this more directly? In the end I consider the expectation of $c_3$ under two different assumptions.. I could compute this directly, or perhaps it is more fitting to consider some variant of the bayes factor here?*

*The disadvantage of the approach with auto-encorders is that generally unfamiliar data is labeled as anomalies.*

## 10.6 ANALYSIS

*Each AE ($AE_{obs}$ and $AE_{bar}$) are trained for* 5000 *iterations on the rotation curve data from the* 152 *mock galaxies. Figure* **??** *in appendix* **??** *show the Hamiltonian for each AE as a function of iterations. As is evident figure* **??***, convergence to the stationary distribution is achieved around iteration* $\approx 950$*. To be conservative, the first* 1000 *iterations of each training is treated as burn in. To measure the accuracy with which the AE's can reproduce the input the average residual squared per output, $\zeta$, is considered*

$$\zeta^{(i)} \equiv \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_{p_8}}\sum_{q\in p_8}\xi_{k,q}(f_k(x^{(i)},\theta_q) - y_k^{(i)})\right)^2, \quad (243)$$

*with* $N_{p_8} = 4000$ *(i.e. the* 5000 *iterations minus burn in).* $\zeta^{(i)}$ *is a measure similar to the chi square per test data point. Figure* **??** *in appendix* **??** *show* $\zeta^{(i)}$ *for the test data. The goal is to identify instances for which the input is reproduced less well and analyze the correspondence between $AE_{obs}$ and $AE_{bar}$ for these instances. For this reason, an anomaly is defined as $\zeta^{(i)} > A$, with $A$ being a threshold. Associated with the anomalies three different counters of anomalies are defined*

1. *$C_1$ the number of instances, $i$, for which an anomaly is only detected in $AE_{bar}$ (i.e. $\zeta^{(i)}_{AE_{obs}} < A \wedge \zeta^{(i)}_{AE_{bar}} > A$).*

2. *$C_2$ the number of instances, $i$, for which an anomaly is only detected in $AE_{obs}$ (i.e. $\zeta^{(i)}_{AE_{obs}} > A \wedge \zeta^{(i)}_{AE_{bar}} < A$).*

3. *$C_3$ the number of instances, $i$, for which an anomaly is detected in both $AE_{bar}$ and $AE_{obs}$ (i.e. $\zeta^{(i)}_{AE_{obs}} > A \wedge \zeta^{(i)}_{AE_{bar}} > A$).*

*From the definitions of $C_1, C_2$ and $C_3$, the fraction of anomalies corresponding to $C_1, C_2$ and $C_3$ are defined viz*

$$\begin{aligned} c_1 &\equiv \frac{C_1}{C_1 + C_2 + C_3}, \\ c_2 &\equiv \frac{C_2}{C_1 + C_2 + C_3}, \\ c_3 &\equiv \frac{C_3}{C_1 + C_2 + C_3}. \end{aligned} \tag{244}$$

*In relation to Renzo's rule, $c_3$ – the fraction of anomalies that appear in both the baryonic and observed rotation curves – is of particular interest, since in the case where Renzo's rule is completely accurate and the anomalies detected by the AE's correspond to features in the baryonic and observed rotation curves, $c_3 = 1$ and $c_1 = c_2 = 0$. Figure 11 show the measured $c_1, c_2$ and $c_3$ as function of $A$ for the test data (SPARC galaxies). A small sanity check is that when $A = 0$ everything is counted as anomalies and hence $c_3 = 1$ by definition. Because $\zeta$ can be interpreted as a measure similar to chi-square per degree of freedom $A = 1$ loosely translates to counting the points that are anomalous with more than 1 sigma. For $A \simeq 1$*

$$c_1 \approx 0.2, \qquad c_2 \approx 0.4, \qquad c_3 \approx 0.4. \tag{245}$$

*In order to interpret the values of $c_3$ from figure 11, it is expedient to consider a rough reference point; the expected value of $c_3$ can roughly be represented as*

$$\mathbb{E}[c_3] \approx \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{\mathbb{P}_1^{(i)}}{\mathbb{P}_2^{(i)}}. \tag{246}$$

*with $N_{test} = 3138$ being the number of test points, $\mathbb{P}_1^{(i)}$ the probability of simultaneous anomalies and $\mathbb{P}_2^{(i)}$ the probability of an anomaly. Formally*

$$\begin{aligned} \mathbb{P}_1^{(i)} &= \mathbb{P}(\zeta_{obs}^{(i)} > A, \zeta_{bar}^{(i)} > A | x^{(i)}, \xi, \theta, I), \\ \mathbb{P}_2^{(i)} &= \mathbb{P}_3^{(i)} + \mathbb{P}_4^{(i)} - \mathbb{P}_1^{(i)}, \end{aligned} \tag{247}$$

*with*

$$\begin{aligned} \mathbb{P}_3^{(i)} &\equiv \mathbb{P}(\zeta_{obs}^{(i)} > A | x^{(i)}, \xi, \theta, I), \\ \mathbb{P}_4^{(i)} &\equiv \mathbb{P}(\zeta_{bar}^{(i)} > A | x^{(i)}, \xi, \theta, I). \end{aligned} \tag{248}$$

*If the anomalies in the baryonic and observed rotation curves are independent then*

$$\mathbb{P}_1^{(i)} = \mathbb{P}_3^{(i)} \mathbb{P}_4^{(i)}. \tag{249}$$

*Assuming for simplicity that there is no correlation between different anomalous data points in the observed rotation curve and different anomalous data points in the baryonic rotation curve, then*

$$\mathbb{P}_3^{(i)} \approx \frac{n_{ano}^{(obs)}}{N_{test}}, \qquad \mathbb{P}_4^{(i)} \approx \frac{n_{ano}^{(bar)}}{N_{test}}. \tag{250}$$

*with $n_{ano}^{(obs)}$ being the number of test points for which $\zeta_{obs}^{(i)} > A$ and $n_{ano}^{(bar)}$ being the number of test points for which $\zeta_{bar}^{(i)} > A$. Combining equations* (246)-(250) *yield*

$$\mathbb{E}[c_3] \approx \frac{n_{ano}^{(obs)} n_{ano}^{(bar)}}{N_{test} \left( n_{ano}^{(obs)} + n_{ano}^{(obs)} - \frac{n_{ano}^{(obs)} n_{ano}^{(bar)}}{N_{test}} \right)}. \tag{251}$$

*The result is displayed as the black line in figure 11. In the figure the green line show the fraction of anomalous points that are coincident, $c_3$, in the two AE's as a function of the threshold used to define an anomaly, $A$. As stated previously, in an idealized setting with perfect data and method, Renzo's rule predicts $c_3 = 1$ for all $A$. However, because of i) measurement imperfections in the independent measurements techniques of the baryonic and observed rotation curves and ii) a non-perfect correspondence between anomalies detected by the BANNs and features that may be subjectively defined (see appendix* **??** *for an extended discussion on this topic), $0 < c_3 < 1$ is expected for $A > 0$ even if Renzo's*

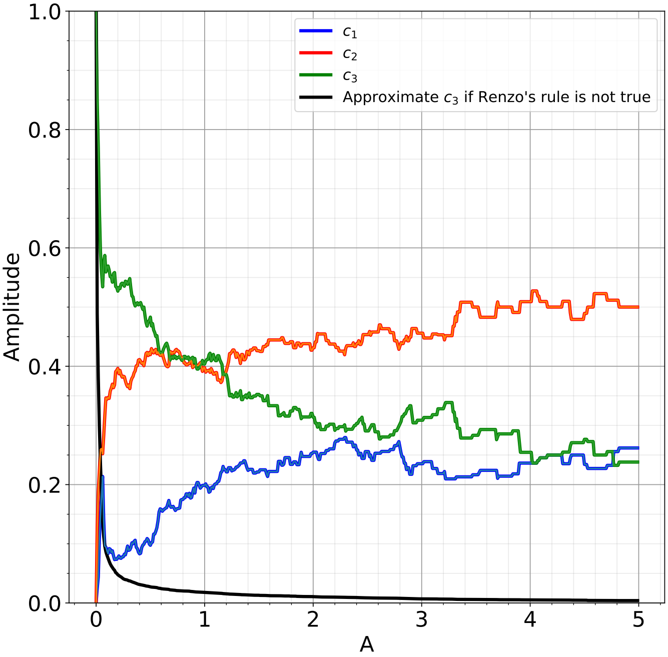

Figure 11: $c_1$ (blue), $c_2$ (orange) and $c_3$ (green) as a function of the threshold for detection, $A$, for the test data (SPARC galaxies). Reproduction of figure 1 from Paper VI [**Petersen:2020renzo**].

*rule is completely accurate. The black curve in figure 11 show the value of $c_3$ that is naively expected if anomalies in the baryonic and observed rotation curves are random and hence Renzo's rule is not true. Comparing the green and black curves from figure 11, it is clear that – although it is not possible to determine an exact expected value for $c_3$ in case Renzo's rule is true – within the simplified setting considered, correlations between anomalies in the baryonic and observed rotation curves that are significantly higher than random are found– as required by Renzo's rule. When* *an anomaly is present in either the baryonic or observed rotation curve, the anomaly is coincident (i.e. an anomaly is detected in both rotation curves at the same instance) in about* 40% *of instances. Figures* **??**-**??** *in appendix* **??** *show rotation curves for all the galaxies with detected anomalies using* $A = 1$. *Reviewing the detected anomalies it is clear that coincident anomalies are abundant at all radii.*

**Example 10.6.**

Consider a discrete action space with an observation $X = x$ and available data $D$. Picking a class corresponds to an action, so classification can be viewed as a game against nature, where nature has picked the true class and the robot has to pick a class as well. Suppose there are only two classes and the cost function is defined by the matrix

$$\begin{array}{cc|cc|}
 & & \multicolumn{2}{c}{S} \\
 & & s_1 & s_2 \\
\cline{3-4}
U(x) & u_1 & 0 & \lambda_{01} \\
 & u_2 & \lambda_{10} & o \\
\cline{3-4}
\end{array}$$

*1.* Show that the decision $u$ that minimizes the expected loss is equivalent to setting a probability threshold $\theta$ and predicting $U(x) = u_1$ if $p(S = s_1 | x, D, I) < \theta$ and $U(x) = u_2$ if $p(S = s_2 | x, D, I) \geq \theta$. What is $\theta$ as a function of $\lambda_{01}$ and $\lambda_{10}$?

*The conditional expected cost*

$$\begin{aligned}
\mathbb{E}_{S|X}[C(u, S)|x, D, I] &= \sum_s C(u, S = s) p(S = s | x, D, I) \\
&= C(u, S = s_1) p(S = s_1 | x, D, I) \\
&\quad + C(u, S = s_2) p(S = s_2 | x, D, I)
\end{aligned} \tag{252}$$

*For the different possible actions*

$$\begin{aligned}
\mathbb{E}_{S|X}[C(u_1, S)|x, D, I] &= \lambda_{01} p(S = s_2 | x, D, I), \\
\mathbb{E}_{S|X}[C(u_2, S)|x, D, I] &= \lambda_{10} p(S = s_1 | x, D, I),
\end{aligned} \tag{253}$$

$U(x) = u_1$ *iff*

$$\mathbb{E}_{S|X}[C(u_1, S)|x, D, I] < \mathbb{E}_{S|X}[C(u_1, S)|x, D, I]) \tag{254}$$

*meaning*

$$\begin{aligned}\lambda_{01}p(S=s_2|x,D,I) &< \lambda_{10}p(S=s_1|x,D,I)\\ &= \lambda_{10}(1-p(S=s_2|x,D,I))\end{aligned} \tag{255}$$

*meaning* $U(x)=u_0$ *iff*

$$p(S=s_2|x,D,I) < \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}} = \theta \tag{256}$$

2. Show a loss matrix where the threshold is 0.1.

$\theta = \frac{1}{10} = \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}} \Rightarrow \lambda_{01} = 9\lambda_{10}$ *yielding the loss matrix*

$$\begin{array}{cc|cc|} & & & S \\ & & s_1 & s_2 \\ \cline{3-4} U(x) & u_1 & 0 & 9\lambda_{10} \\ & u_2 & \lambda_{10} & o \\ \cline{3-4} \end{array}$$

*You may set* $\lambda_{10}=1$ *since only the relative magnitude is important in relation to making a decision.*

---

**Example 10.7.**

In many classification problems one has the option of assigning $x$ to class $k \in K$ or, if the robot is too uncertain, choosing a reject option. If the cost for rejection is less than the cost of falsely classifying the object, it may be the optimal action. Define the cost function as follows

$$C(u,s) = \begin{cases} 0 & \textit{if correct classification } (u=s) \\ \lambda_r & \textit{if reject option } u = \textit{reject} \\ \lambda_s & \textit{if wrong classification } (u \neq s) \end{cases} . \tag{257}$$

1. Show that the minimum cost is obtained if the robot decides on class $u$ if $p(S=u|x,D,I) \geq p(S \neq u|x,D,I)$ and if $p(S=u|x,D,I) \geq 1-\frac{\lambda_r}{\lambda_s}$.

*The conditional expected cost if the robot does not pick the reject option, meaning $u \in \mathbb{U} \setminus$ reject*

$$\begin{aligned}\mathbb{E}_{S|X}[C(u,S)|x,D,I] &= \sum_s C(u,S=s)p(S=s|x,D,I) \\ &= \sum_{s \neq u} \lambda_s p(S=s|x,D,I) \\ &= \lambda_s(1-p(S=u|x,D,I))\end{aligned} \tag{258}$$

*where for the second equality it has been used that the cost of a correct classification is* $0$*, so the case of* $S=u$ *does not enter the sum. For the third equality it has been used that summing over all but* $S=u$ *is equal to* $1-p(S=u|x,D,I)$*. The larger* $p(S=u|x,D,I)$*, the smaller loss (for* $\lambda_s>0$*), meaning the loss is minimized for the largest probability. The conditional expected loss if the robot picks the reject option*

$$\begin{aligned}\mathbb{E}_{S|X}[C(reject,S)|x,D,I] &= \lambda_r \sum_s p(S=s|x,D,I) \\ &= \lambda_r.\end{aligned} \tag{259}$$

*Equation* (258) *show picking* $\arg\max_{u \in \mathbb{U} \setminus reject} p(S=u|x,D,I)$ *is the best option among classes* $u \neq$ *reject. To be the best option overall, it also needs to have lower cost than the reject option. Using equations* (258) *and* (259) *yields*

$$(1-p(S=u|x,D,I))\lambda_s < \lambda_r \tag{260}$$

*meaning*

$$p(S=u|x,D,I) \geq 1-\frac{\lambda_r}{\lambda_s}. \tag{261}$$

2. Describe qualitatively what happens as $\frac{\lambda_r}{\lambda_s}$ is increased from 0 to 1.

$\frac{\lambda_r}{\lambda_s} = 0$ *means rejection is rated as a successful classification – i.e. no cost associated – and this become the best option (rejection that is) unless* $p(y = j|x) = 1$*, corresponding to knowing the correct class with absolute certainty. In other words; in this limit rejection is best unless the robot is certain of the correct class.* $\frac{\lambda_r}{\lambda_s} = 1$ *means rejection is rated a misclassification – i.e.* $\lambda_r = \lambda_s$ *– and thus and "automatic cost". Hence, in this case rejection is never chosen. In between the limits, an interpolation of interpretations apply.*

**Example 10.8.**

Consider a 3-class naive Bayes classifier[3] with one binary feature and one Gaussian feature $y \sim \text{Mu}(y|\pi, 1)$, $x_1|y = c \sim \text{Ber}(x_1|\theta_c)$, $x_2|y = c \sim N(x_2|\mu_c, \sigma_c^2)$ with

$$\pi = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \end{pmatrix}, \theta = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \mu = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \sigma^2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{262}$$

1. Compute $p(y|x_1 = 0, x_2 = 0)$.

$$\begin{aligned} p(y = c|x_1 = 0, x_2 = 0, I) &= \frac{p(x_1 = 0, x_2 = 0|y = c, I)p(y = c|I)}{p(x_1 = 0, x_2 = 0|I)} \\ &= \frac{p(x_1 = 0|x_2 = 0, y = c, I)p(x_2 = 0|y = c, I)p(y}{p(x_1 = 0, x_2 = 0|I)} \\ &= \frac{p(x_1 = 0|y = c, I)p(x_2 = 0|y = c, I)p(y = c|I)}{p(x_1 = 0, x_2 = 0|I)} \\ &= \frac{p(x_1 = 0|y = c, I)p(x_2 = 0|y = c, I)p(y = c|}{\sum_{c'} p(x_1 = 0|y = c', I)p(x_2 = 0|y = c', I)p(y =} \end{aligned}$$

[3] The naive Bayes classifier assume the class conditional density can be written as a product of one-dimensional densities [39, p.84]

$$\tag{263}$$

*with*

$$\begin{aligned} p(x_1 = 0|y, I) &= Ber(x_1 = 0|\theta) \\ &= \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \\ p(x_2 = 0|y, I) &= N(x_2 = 0|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x_2-\mu}{\sigma})^2} \\ &\simeq \begin{pmatrix} 0.24 \\ 0.4 \\ 0.24 \end{pmatrix} \\ p(y|\pi) &= \pi \end{aligned} \tag{264}$$

$$p(y = c|x_1 = 0, x_2 = 0, I) = \begin{pmatrix} 0.43 \\ 0.35 \\ 0.22 \end{pmatrix} \tag{265}$$

2. Compute $p(y|x_1 = 0, I)$.

$$\begin{aligned} p(y = c|x_1 = 0, I) &= \frac{p(x_1 = 0|y = c, I)p(y = c|I)}{p(x_1 = 0|I)} \\ &= \frac{p(x_1 = 0|y = c, I)p(y = c|I)}{\sum_{c'} p(x_1 = 0|y = c', I)p(y = c'|I)} \\ &= \pi \\ &= \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \end{pmatrix} \end{aligned} \tag{266}$$

3. Compute $p(y|x_2 = 0, I)$.

$$\begin{aligned} p(y = c|x_2 = 0, I) &= \frac{p(x_2 = 0|y = c, I)p(y = c|I)}{p(x_2 = 0|I)} \\ &= \frac{p(x_2 = 0|y = c, I)p(y = c|I)}{\sum_{c'} p(x_2 = 0|y = c', I)p(y = c'|I)} \\ &= \begin{pmatrix} 0.43 \\ 0.35 \\ 0.22 \end{pmatrix} \end{aligned} \tag{267}$$

4. Explain any interesting patterns you see in your results.

*Since $\theta_c = \theta_j \forall c, j \in \{c\}$ the prior information is propagated through in calculating $p(y|x_1 = 0, I)$ and $x_1$ does not impact $p(y|x_1 = 0, x_2 = 0, I)$, as can be seen by comparing $p(y|x_1 = 0, x_2 = 0, I)$ and $p(y|x_2 = 0, I)$. This can be seen by considering*

$$\begin{aligned} p(y = c|x_1 = 0, x_2 = 0, I) &= \frac{p(x_1 = 0|y = c, I)p(x_2 = 0|y = c, I)p(y = c|}{\sum_{c'} p(x_1 = 0|y = c', I)p(x_2 = 0|y = c', I)p(y =} \\ &= \frac{p(x_2 = 0|y = c, I)p(y = c|I)}{\sum_{c'} p(x_2 = 0|y = c', I)p(y = c'|I)} \\ &= p(y = c|x_2 = 0, I) \end{aligned} \tag{268}$$

*where it has been used that $p(x_1 = 0|y = c, I) = p(x_1 = 0|y = c', I) \forall c, c' \in \{c\}$ for the second equality.*

---

**Example 10.9.**

*Consider a farmer who wishes to retire and therefore would like to* *sell their set of live animals. For simplicity, assume that animals make up a simple group and a given person can either be interested in purchasing an animal or not – the simplfication consist of not differentiating between different animal types. In order to sell their animals, the farmer needs to contact people with a sale in mind. For simplicity, the contact will be assumed to be via a telephone call only. The farmer can call (or not) with the intent to sell an animal to the recepient of the call. The recepient of the call can (or not) be interesed in purchasing an animal (Natures decision). Let $\mathbb{U}$ denote the set of the farmers actions and $\mathbb{S}$ the set of Natures actions, then*

$$\begin{aligned} \mathbb{U} &= \{u_1 = \textit{call}, u_2 = \textit{don't call}\}, \\ \mathbb{S} &= \{s_1 = \textit{interested}, s_2 = \textit{not interested}\}. \end{aligned} \tag{269}$$

*A nuisance for the contacted people is associated to the call which is represented by the abstract monetary loss, $\lambda \in \mathbb{R}^+$. The degree to which people are annoyed by a sales call is independent in general and the monetary loss represents the average animosity generated and the associated moneraty loss connected to a worsened reputation. Aside from the nuisance associated with a sales call, there is also a moneraty reward for a successfull sale, $\psi$. If the farmer cannot sell his animals, he will have them terminated with no associated cost in order not to spend additional time or money on them. Given these assumptions, the cost function can be represented by the matrix*

$$\begin{array}{cc|cc|} & & \multicolumn{2}{c}{S} \\ & & s_1 = \textit{Interested} & s_2 = \textit{Not interested} \\ \cline{3-4} U & u_1 = \textit{Call} & \lambda - \psi & \lambda \\ & u_2 = \textit{Don't call} & 0 & o \\ \cline{3-4} \end{array}$$

*The farmer has available to them observations $X = x$ that contain information regarding the decesion Nature is going to make, $S = s$. The farmer also have a collection of past observations and resulting decisions of Nature, i.e. $D = \{(X = x_1, S = s_1), (X =$*

*$x_2, S = s_2), \ldots (X = x_n, S = s_n)\} = D_s \times D_s$. The optimal decision for the farmer to call the i'th person can then be written viz*

$$U^*(x) = \arg\min_{U(x)} \mathbb{E}_{S|X}[C(U(x), S)|X = x, D, I], \qquad (270)$$

*where*

$$\mathbb{E}_{S|X}[C(U(x), S)|X = x, D, I] = \sum_{s \in \mathbb{S}} C(U(x), s) p(S = s|X = x, D, I). \qquad (271)$$

*Writing out the conditional expectation*

$$\begin{aligned}
\mathbb{E}_{S|X}[C(u_1, S)|x, D, I] &= \sum_s C(u_1, s) p(S = s|x, D, I) \\
&= C(u_1, s_1) p(S = s_1|x, D, I) + C(u_1, s_2) p(S = s_2|x, D, \\
&= (\lambda - \psi) p(S = s_1|x, D, I) + \lambda_i p(S = s_2|x, D, I) \\
\mathbb{E}_{S|X}[C(u_2, S)|x, D, I] &= \sum_s C(u_2, s) p(S = s|x, D, I) \\
&= C(u_2, s_1) p(S = s_1|x, D, I) + C(u_2, s_2) p(S = s_2|x, D, \\
&= 0
\end{aligned} \qquad (272)$$

*The optimal decision rule $U^*(x)$ can be implicitly specified as picking $u_1$ (call) iff $\mathbb{E}_{S|X}[C(u_1, S)|x, D, I] < \mathbb{E}_{S|X}[C(u_2, S)|x, D, I]$, corresponding to picking $u_1$ (call) iff*

$$(\lambda - \psi) p(S = s_1|x, D, I) + \lambda p(S = s_2|x, D, I) < 0 \qquad (273)$$

*Since $p(S = s_1|x, D, I) + p(S = s_2|x, D, I) = 1$*

$$\frac{\lambda}{\psi} < p(S = s_1|x, D, I) \qquad (274)$$

*meaning the farmer should call (action $u_1$) iff the probability for the recepient of the call to be interested in at least one animal is larger than the penalty of calling divided by the gain of calling.*

### 10.6.1 *The Probability*

*Equation* (274) *implicitly specify the decision rule for the farmer. $\lambda, \psi$ is assumed specified, so only the probability $p(S = s_1|x, D, I)$ remain to be specified. Using marginalization and assuming independence*

$$\begin{aligned} p(S = s|x, D, I) &= \int p(S = s, \theta|x, D, I)d\theta \\ &= \int p(S = s|x, \theta, D, I)p(\theta|x, D, I)d\theta \\ &= \int p(S = s|x, \theta, I)p(\theta|D, I)d\theta. \end{aligned} \tag{275}$$

*Suppose now a model, $f : \mathbb{W} \times X \mapsto [0, 1]$, with associated parameters $\theta \in \mathbb{W}$, that estimates Nature's actions $S$ based on observed data $X$ is introduced. The random variable $S$ is discrete and the function is identified with the probability of each action*

$$p(S = s|x, \theta, I) = f_{S=s}(\theta, x), \tag{276}$$

*with*

$$\sum_{s \in \mathbb{S}} p(S = s|x, \theta, I) = 1. \tag{277}$$

*Combining equation* (275) *and* (276)

$$\begin{aligned} p(S = s|x, D, I) &= \int f_{S=s}(\theta, x)p(\theta|D, I)d\theta \\ &= \mathbb{E}[f_{S=s}(\theta, x)|D, I]. \end{aligned} \tag{278}$$

*From Bayes theorem*

$$p(\theta|D, I) = \frac{p(D_s|D_x, \theta, I)p(\theta|D_x, I)}{p(D_s|D_x, I)}, \tag{279}$$

*where $p(\theta|D_x, I) = p(\theta|I)$. Assuming the distribution over $\theta$ is normally distributed with zero mean and a precision described by a hyperparameter, $\lambda$,*

$$p(\theta|I) = \int p(\theta|\lambda, I)p(\lambda|I)d\lambda. \tag{280}$$

*The precision is constructed as a wide gamma distribution so as to approximate an objective prior*

$$p(\theta|\lambda,I)p(\lambda|I)=\prod_{q=1}^{\tilde{n}}\frac{\lambda_q^{\frac{n_q}{2}}}{(2\pi)^{\frac{n_q}{2}}}e^{-\frac{\lambda_q}{2}\sum_{l=1}^{n_q}\theta_l^2}\frac{\beta_q^{\alpha_q}}{\Gamma(\alpha_q)}\lambda_q^{\alpha_q-1}e^{-\beta_q\lambda_q} \tag{281}$$

*Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written*

$$\begin{aligned} p(D_s|D_x,\theta,I) &= \prod_{i=1}^{n} p(S=s_i|X=x_i,\theta,I) \\ &= \prod_{i=1}^{n} f_{s_i}(\theta,x_i) \end{aligned} . \tag{282}$$

*Aside from the specification of the model $f$, $p(S=s|x,D,I)$ is at this point fully specified and can be approximated by HMC similarly to the regression case. In this case, the model can be represented by the Hamiltonian*

$$H\equiv\sum_q\sum_l\frac{p_l^2}{2m_l}-\ln(p(\theta,\lambda|D,I))+const \tag{283}$$

*where*

$$p(\theta|D,I)=\int d\lambda p(\theta,\lambda|D,I). \tag{284}$$

*Using equations* (275)-(282) *in equation* (283) *yields the Hamiltonian*

$$\begin{aligned} H=&\sum_{q=1}^{\tilde{n}}\sum_{l=1}^{n_q}\frac{p_l^2}{2m_l}-\sum_{i=1}^{n}\ln(f_{s_i}(\theta,x_i))+const \\ &+\sum_{q=1}^{\tilde{n}}\Bigg(\ln(\Gamma(\alpha_q))-\alpha_q\ln(\beta_q)+(1-\alpha_q)\ln(\lambda_q)+\beta_q\lambda_q. \\ &\qquad+\frac{n_q}{2}(\ln(2\pi)-\ln(\lambda_q))+\frac{\lambda_q}{2}\sum_{l=1}^{n_q}\theta_l^2\Bigg) \end{aligned} \tag{285}$$

### 10.6.2 *Simple Model*

*Let*

$$f_{S=s}(\theta,x_i)=\frac{e^{b_s+\sum_q a_{sq}x_{iq}}}{\sum_{k\in\mathbb{S}} e^{b_k+\sum_q a_{kq}x_{iq}}}, \tag{286}$$

*where* $\theta=\{b,a\}$.

### 10.6.3 *Manual HMC Algorithm*

*The Hamiltonian is given by*

$$\begin{aligned} H=&\sum_{q=1}^{2}\sum_{l=1}^{2}\frac{p_{ql}^2}{2m_{ql}}-\sum_{i=1}^{n}\ln(f_{s_i}(\theta,x_i))\\ &+\ln(\Gamma(\alpha_a))-\alpha_a\ln(\beta_a)+(1-\alpha_a)\ln(\lambda_a)+\beta_a\lambda_a\\ &\qquad+\frac{1}{2}(\ln(2\pi)-\ln(\lambda_a))+\frac{\lambda_a}{2}\sum_{j,q}a_{jq}^2\\ &+\ln(\Gamma(\alpha_b))-\alpha_b\ln(\beta_b)+(1-\alpha_b)\ln(\lambda_b)+\beta_b\lambda_b\\ &\qquad+\frac{1}{2}(\ln(2\pi)-\ln(\lambda_b))+\frac{\lambda_b}{2}\sum_{j}b_j^2 \end{aligned} \tag{287}$$

$\lambda_j$ *is positive definite. In order to uphold this numerically, let* $\lambda_j=e^{\tau_j}$. *When making this transformation, the integration measure of equation* (280) *has to be transformed as well. This proceeds viz*

$$d\lambda_j=\lambda_j d\tau_j, \tag{288}$$

*meaning effectively* $\lambda_j$ *is multiplied on* $p(\theta,\lambda|D,I)$ *such that* $H\to H-\ln(\lambda_j)$. *This means*

$$(1-\alpha_j)\ln(\lambda_j)\in H\Rightarrow -\alpha_j\ln(\lambda_j). \tag{289}$$

*Additionally, it is convenielt to pick out the $s_i$ via a one-hot target vector such that*

$$\begin{aligned} H = & \sum_{q=1}^{2}\sum_{l=1}^{2}\frac{p_{ql}^2}{2m_{ql}} - \sum_{j\in\mathbb{S}}\sum_{i=1}^{n} s_{ij}\ln(f_j(\theta, x_i)) \\ & + \ln(\Gamma(\alpha_a)) - \alpha_a\ln(\beta_a) - \alpha_a\tau_a + \beta_a e^{\tau_a} \\ & \qquad + \frac{1}{2}(\ln(2\pi) - \tau_a) + \frac{e^{\tau_a}}{2}\sum_{j,q} a_{jq}^2 \\ & + \ln(\Gamma(\alpha_b)) - \alpha_b\ln(\beta_b) - \alpha_b\tau_b + \beta_b e^{\tau_b} \\ & \qquad + \frac{1}{2}(\ln(2\pi) - \tau_b) + \frac{e^{\tau_b}}{2}\sum_{j} b_j^2 \end{aligned} \tag{290}$$

*The derivatives are needed for the HMC algorithm*

$$\frac{\partial H}{\partial a_{ml}} = -\sum_{i,j}\frac{s_{ij}}{f_{ij}}\frac{\partial f_{ij}}{\partial a_{ml}} + e^{\tau_a}a_{ml}, \tag{291}$$

$$\begin{aligned} \frac{\partial f_{ij}}{\partial a_{ml}} = & \frac{e^{b_j+\sum_{q_1} a_{jq_1}x_{iq_1}}}{\sum_{k\in\mathbb{S}} e^{b_k+\sum_{q_2} a_{kq_2}x_{iq_2}}}\sum_{q_3}\delta_{jm}\delta_{q_3l}x_{iq_3} \\ & - \frac{e^{b_j+\sum_{q_4} a_{jq_4}x_{iq_4}}}{(\sum_{k\in\mathbb{S}} e^{b_k+\sum_{q_5} a_{kq_5}x_{iq_5}})^2}\sum_{k'\in\mathbb{S}} e^{b_{k'}+\sum_{q_6} a_{k'q_6}x_{iq_6}}\sum_{q_7}\delta_{k'm}\delta_{q_7l}x_{iq_7} \\ = & f_{ij}\delta_{jm}x_{il} - f_{ij}f_{im}x_{il} \end{aligned} \tag{292}$$

*where it has been used that*

$$\frac{\partial a_{jq_3}}{\partial a_{ml}} = \delta_{jm}\delta_{q_3l} \tag{293}$$

$$\begin{aligned} \frac{\partial H}{\partial a_{ml}} & = -\sum_{i,j}\frac{s_{ij}}{f_{ij}}(f_{ij}\delta_{jm}x_{il} - f_{ij}f_{im}x_{il}) + e^{\tau_a}a_{ml} \\ & = \sum_{i} x_{il}(f_{im} - s_{im}) + e^{\tau_a}a_m \end{aligned} \tag{294}$$

$$\frac{\partial H}{\partial b_m} = \sum_i (f_{im} - s_{im}) + e^{\tau_b} b_m. \tag{295}$$

$$\frac{\partial H}{\partial \tau_m} = -\alpha_m + \beta_m e^{\tau_m} - \frac{1}{2} + \frac{e^{\tau_m}}{2} \sum_j m_j^2. \tag{296}$$

*The masses for the HMC algorithm can be set by approximating the second order derivatives as fixed. Let*

$$\begin{aligned} \frac{\partial^2 H}{\partial a_{ml}^2} &= \sum_i x_{il} \frac{\partial f_{im}}{\partial a_{ml}} + e^{\tau_a} \\ &= \sum_i x_{il}^2 f_{im}(1 - f_{im}) + e^{\tau_a} \end{aligned} \tag{297}$$

*then taking* $x_{il}^2 \sim 1, f_{im} \sim \frac{1}{2}$ *and any parameter* $\sim 0$*, meaning* $e^{\tau_a} \sim 1$ *yield the mass approximation*

$$\begin{aligned} m_{ml}^{(a)} &\sim \left.\frac{\partial^2 H}{\partial a_{ml}^2}\right|_{\textit{fixed approximation}} \\ &\sim N \cdot 1^2 \cdot \frac{1}{2}(1 - \frac{1}{2}) + 1 \\ &= \frac{N}{4} + 1, \end{aligned} \tag{298}$$

*where N is the number of data samples in* $D_x$*. Similarly*

$$\begin{aligned} \frac{\partial^2 H}{\partial b_m^2} &= \sum_i \frac{\partial f_{im}}{\partial a_{ml}} + e^{\tau_b} \\ &= \sum_i f_{im}(1 - f_{im}) + e^{\tau_b}, \end{aligned} \tag{299}$$

*meaning (since* $x_{il}^2 \sim 1$*)*

$$\begin{aligned} m_{ml}^{(a)} &\sim \left.\frac{\partial^2 H}{\partial b_m^2}\right|_{\textit{fixed approximation}} \\ & m_{ml}^{(b)}. \end{aligned} \tag{300}$$

*The precision parameter*

$$\frac{\partial^2 H}{\partial \tau_q^2} = \beta_q e^{\tau_q} + \frac{e^{\tau_q}}{2} \sum_j q_j^2. \tag{301}$$

*Take* $\beta_q = 3$*, then*

$$\begin{aligned} m_q^{(\tau)} &\sim \left.\frac{\partial^2 H}{\partial \tau_q^2}\right|_{\textit{fixed approximation}} \\ &\sim 3. \end{aligned} \tag{302}$$

### 10.6.4 *Data*

*Take* $x = (\textit{area}, \textit{number of animals})^T$ *and* $s =$ *2d one-hot vector, with* $\dim(D) = 1000$*.* $D$ *is split into two sets* $D^{(\textit{training})}$ *and* $D^{(\textit{test})}$*, with* $\dim(D^{(\textit{training})}) = \gamma \dim(D)$ *and* $\dim(D^{(\textit{test})}) = (1-\gamma)\dim(D)$ *and* $\gamma = 0.6$*.* $D^{(\textit{training})}$ *will be used to train the model and* $D^{(\textit{test})}$ *to evaluate the quality of the trained model. The underlying truth of Nature (unbeknownst to the model) is that an animal will be purchased iff*

$$\textit{total area} - 3.2 \cdot \textit{number of animals} \geq 3.2. \tag{303}$$

### 10.6.5 *Training*

*Using* $D^{(\textit{training})}$ *as input, the algorithms the algorithms are trained for* 2000 *iterations. The first* 500 *iterations are taken as burn in to be conservative. The coefficients,* $\theta$*, for iterations* $[500, 2000]$ *are used to make a model prediction viz*

$$p(S = s|x, D, I) = \frac{1}{1500} \sum_{i=500}^{2000} f(\theta_i, x) \tag{304}$$

*The accuracy of the modeled probabilities can be gauged by considering the case where* $\psi = 2\lambda$ *such that the decision rule (equation* (274)*) becomes*

$$\frac{1}{2} < p(S = s_1|x, D, I), \tag{305}$$

*and the classification is driven by the probabilities alone.*

### 10.6.6 *PyMC HMC Algorithm*

*PyMC is a probabilistic programming library for Python that allows users to build Bayesian models with a Python API and fit them using Markov Chain Monte Carlo methods PyMC link. Using this API it is possible to create, train and test a PyMC-equivalent of the model described in the previous sections. The general approach to building a model using PyMC consists of stating the data generating process, specifying a likelihood, and related prior distributions for any parameters involved. While modeling the data generating process and framing the statistical problem correctly are never completely trivial and require some effort from the user it is rather straight forward to perform the Markov Chain Monte Carlo sampling with PyMC. The user do not need to do any calculations related to the Hamiltonian Monte Carlo method nor do they need to specifically handle any integrals. In addition, there is a range of predefined probability distributions both discrete and continuous readily available in the library such as the Gamma (figure 12) and Normal distribution used for modeling the priors as described in section 10.6.1.*

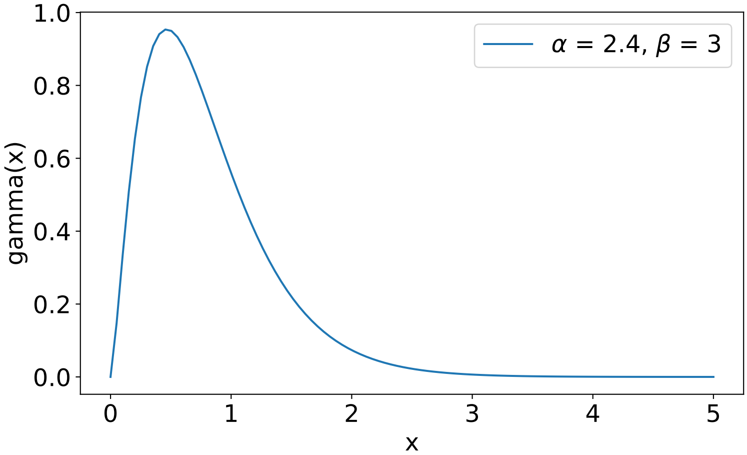

Figure 12: Plot of Gamma probability distribution used as priors for the precision parameters in Normal distributions.

*The python code in algorithm* **??** *shows how the simple model can be declared using PyMC. All models in PyMC are declared using a "with pm.Model() as modelname"-statement, where pm is the abbreviated form of the PyMC package. Within this statement at range of random variables, data and their relations can be declared. The syntax for declaring a random variable by its probability distribution (such as pm.Normal) is to pass a name for the random variable as a string for the first argument. The rest of the arguments are typically values for the parameters specific to the distribution. PyMC random variables can often be used for stating parameters in other random variables forming a hierarchy of distributions as is the case for the precision variables used to model the variance for the normally distributed parameters a and b (see equation 286) with zero mean. In the code example, these parameters are then combined deterministically with $D_x$ to constitute the data generating process of the simple model. $D_x$ is declared as a PyMC mutable object so that it will be possible later to change the values to generate predictions. As a last step the likelihood is declared. In PyMC this is the step where the relationship between* *the observed conditional ($D_s$) data and parameters is stated. As the output is the probabilities for the two classes a Multinomial distribution with the number of independent trials (n) equal to 1 is used as the likelihood.*

**Algorithm 1** PyMC Python Code

```
import pymc as pm
import numpy as np
with pm.Model() as classifier_model:
covars = pm.MutableData(’covars’,data_x_training)
# Priors for precision
precision_a = pm.Gamma(’precision_a’, alpha=2.4, beta=3)
precision_b = pm.Gamma(’precision_b’, alpha=2.4, beta=3)
# Priors for parameters
param_a = pm.Normal(’param_a’,
0,
sigma=1/np.sqrt(precision_a),
shape=(2,2))
param_b = pm.Normal(’param_b’,
0,
sigma=1/np.sqrt(precision_b),
shape=(2,))
# Data generating process
T = pm.Deterministic(’T’,pm.Math.exp(param_b + pm.math.
                         dot(covars, param_a.T)))
class_conditional_probability = pm.Deterministic(’
                         class_conditional_probability’,
                          T/T.sum(axis=1, keepdims=True)
                         )
# Likelihood
obs = pm.Multinomial(’obs’, n=1, p=
                         class_conditional_probability,
                         observed=data_s_training, shape
                         =class_conditional_probability.
                         shape)
```

*After declaring the model it is possible to sample the posterior by calling the pm.sample() method. Burn-in can be controlled by setting the tune argument. The draw argument determines how* *many samples are being drawn while a number of chains can be run in parallel by setting the chains and cores (computational) arguments.*

**Algorithm 2** PyMC Posterior Python Code

```
with classifier_model:
posterior = pm.sample(tune=512, draws=1024, chains=4,
                         cores=4)
```

*Using the result of drawing samples from the distribution of model parameters (posterior) it is possible to draw from the posterior predictive distribution using the pm.sample_posterior_predictive() method. Without changing the input data this is equivalent to obtaining the result of applying the trained model on* $D_x^{(training)}$*. By changing the input data to* $D_x^{(test)}$ *using the pm.set_data() method it is possible to obtain the posterior predictive distribution.*

**Algorithm 3** PyMC Posterior Python Code

```
with classifier_model:
posterior_predictive = pm.sample_posterior_predictive(
                         posterior)
pm.set_data({'covars': data_x_test})
posterior_predictive_test = pm.
                         sample_posterior_predictive(
                         posterior)
```

*Using 305 for* $D^{(training)}$*, the PyMC model correctly classify* 598 *of* 600 *data points. For* $D^{(test)}$*, the model correctly classify* 398.

### 10.6.7 *Results*

*Manual HMC Algorithm*

*The HMC algorithm have parameters "step_scale" and "number_of_steps_scale", which adjust the overall scale of the step lengths and number of steps in pase space. Ideally, the distance between points should be large, so that step scale should be small (what the step length is divided by should be small) and the number of steps should be large. Numerical stability only exist for step_scale*$\gtrsim 5$ *(given accurate mass estimation) and thus only the number of steps remain as a variable to tune. In this study step_scale*$= 10$ *and number_of_steps_scale*$= 1500$*, where the latter is limited by reasonable computation time (to match approximately the pymc computation time). Given these parameters, the manual HMC algorithm misclassify a single training data point*

$$x^{(training)}_{misclassified\ 1} = \begin{pmatrix} 3.21798365 \\ 0 \end{pmatrix} \tag{306}$$

*and two test data points*

$$x^{(test)}_{misclassified\ 1} = \begin{pmatrix} 6.40501793 \\ 1 \end{pmatrix}, \quad x^{(test)}_{misclassified\ 2} = \begin{pmatrix} 9.60627827 \\ 2 \end{pmatrix}. \tag{307}$$

*With equation* (303) *in mind, it is clear that the misclassifications of equation* (306) *and* (307) *are close to the limit with respect to purchasing an animal.*

*PyMC HMC Algorithm*

*The PyMC HMC Algorithm obtain misclassify two training data points; equation* (306) *and*

$$x^{(training)}_{misclassified\ 2} = \begin{pmatrix} 12.80826256 \\ 3 \end{pmatrix} \tag{308}$$

*and two test data points (equation* (307)*).*

## 10.7 SUMMARY AND DISCUSSION

*It has been shown how decision theory can be used in conjunction with statistics to make theoretically optimal decisions based on a user specified set of preferences (cost function). Using mock data, a "manual HMC algorithm" written by hand and a standard Python "PyMC HMC algorithm" have been compared. The two yield identical results on test data with the manual model yielding marginally better results on training data. Overall the performace is deemed equivalent both in terms of accuracy and computational speed. The manual HMC algorithm require the user to derive the gradients, write the sampling algorithm in Python and tune the algorith, whereas the latter only require a specification of the model via a standardized PyMC interface. Hence, from a user complexity perspective, the PyMC algorithm has a significant advantage.*

**Example 10.10.**

*Let $x = x_1 + x_2 + \dots x_M$ be the sum of future event counts and $y = \{y_1, y_2, y_3, \dots y_N\}$ be observed event counts. With this notation the expected number of future events, x, and the associated uncertainty can be written*

$$\mathbb{E}[x|y, I] \pm \sqrt{Var[x|y, I]}, \tag{309}$$

*where*

$$\begin{aligned} \mathbb{E}[x|y, I] &= \sum_i ip(x = i|y, I), \\ Var[x|y, I] &= \sum_i i^2 p(x = i|y, I) - \mathbb{E}[x|y, I]^2. \end{aligned} \tag{310}$$

*The distribution over counts $p(x = i|y, I)$ can be expanded by marginalizing over a set of unknown parameters from underlying distributions. The details depend on the statistical assumptions imposed. In this example, different assumptions and their consequences will be investigated.*

SIMPLE POISSON ASSUMPTION: *The most common assumption is to assume data follow a Poisson distribution with an unknown rate parameter which will then be marginalized over. An abvious choice of prior would be the conjugate gamma distribution (which is also the distribution with maximum entropy) with parameters $\alpha, \beta$, meaning*

$$\begin{aligned} p(x = i|y, \alpha, \beta, I) &= \int d\lambda p(x = i, \lambda|y, \alpha, \beta, I) \\ &= \int d\lambda p(x = i|\lambda, \alpha, \beta, y, I) p(\lambda|\alpha, \beta, y, I) \\ &= \int d\lambda p(x = i|\lambda, \alpha, \beta, y, I) \frac{p(y|\lambda, \alpha, \beta, I) p(\lambda|\alpha, \beta, I)}{p(y|\alpha, \beta, I)}. \end{aligned} \tag{311}$$

*with*

$$\begin{aligned}
p(y|\alpha,\beta,I) &= \int d\lambda p(y|\alpha,\beta,\lambda,I)p(\lambda|\alpha,\beta,I),\\
p(y|\alpha,\beta,\lambda,I) &= \prod_{j=1}^{N} Poi(y_j|\lambda)\\
&= \lambda^{N\bar{y}}e^{-N\lambda}\prod_{j=1}^{N}\frac{1}{y_j!},\\
p(\lambda|\alpha,\beta,I) &= Ga(\lambda|\alpha,\beta)\\
&= \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}.
\end{aligned} \tag{312}$$

*Meaning*

$$p(y|\lambda,\alpha,\beta,I)p(\lambda|\alpha,\beta,I) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha+N\bar{y}-1}e^{-(\beta+N)\lambda}\prod_{j=1}^{N}\frac{1}{y_j!} \tag{313}$$

*and consequently*

$$\begin{aligned}
\frac{p(y|\lambda,\alpha,\beta,I)p(\lambda|\alpha,\beta,I)}{p(y|\alpha,\beta,I)} &= \frac{\lambda^{\alpha+N\bar{y}-1}e^{-(\beta+N)\lambda}}{\int d\lambda\lambda^{\alpha+N\bar{y}-1}e^{-(\beta+N)\lambda}}\\
&= Ga(\lambda|\alpha+N\bar{y},\beta+N)
\end{aligned} \tag{314}$$

*Using that* $p(x=i|\lambda,\alpha,\beta,y,I) = Poi(x=i|M\lambda)$ *yields*

$$\begin{aligned}
\mathbb{E}[x|y,\alpha,\beta,I] &= \sum_i i\int d\lambda Poi(x=i|M\lambda)Ga(\lambda|\alpha+N\bar{y},\beta+N)\\
&= M\int d\lambda\lambda Ga(\lambda|\alpha+N\bar{y},\beta+N)\\
&= M\frac{\alpha+N\bar{y}}{\beta+N},
\end{aligned} \tag{315}$$

*where for the second equality it has been used that* $\sum_i iPoi(x=i|M\lambda) = M\lambda$ *and for the second equality it has been realized that* *the second line denoted the expectation of a Gamma distribution with modified parameters. Similarly for the variance*

$$\begin{aligned}\sum_i i^2 p(x=i|y,\alpha,\beta,I) &= \sum_i i^2 \int d\lambda Poi(x=i|M\lambda)Ga(\lambda|\alpha+N\bar{y},\beta+N)\\ &= M\int d\lambda(\lambda+M\lambda^2)Ga(\lambda|\alpha+N\bar{y},\beta+N)\\ &= \mathbb{E}[x|y,\alpha,\beta,I]+M^2\int d\lambda\lambda^2 Ga(\lambda|\alpha+N\bar{y},\beta+N)\\ &= \mathbb{E}[x|y,\alpha,\beta,I]+M^2\frac{\alpha+N\bar{y}}{(\beta+N)^2}+\mathbb{E}[x|y,\alpha,\beta,I]^2\end{aligned} \tag{316}$$

*where for the second equality it has been used that* $\sum_i i^2 Poi(x=i|M\lambda)=Var[i|M\lambda]+\mathbb{E}[x|M\lambda]^2=M\lambda+M^2\lambda^2$ *and similarly for the fourth equality. Combining equations* (310) *and* (316)*, the variance can be written*

$$\begin{aligned}Var[x|y,\alpha,\beta,I] &= \mathbb{E}[x|y,\alpha,\beta,I]+M^2\frac{\alpha+N\bar{y}}{(\beta+N)^2}\\ &= M\left(1+\frac{M}{\beta+N}\right)\frac{\alpha+N\bar{y}}{\beta+N}.\end{aligned} \tag{317}$$

*In the limit of* $\beta,\alpha\to 0$ *equation* (309) *become*

$$\begin{aligned}\mathbb{E}[x|y,I]\pm\sqrt{Var[x|y,I]} &= \lim_{\alpha,\beta\to 0}\left(\mathbb{E}[x|y,\alpha,\beta,I]\pm\sqrt{Var[x|y,\alpha,\beta,I]}\right)\\ &= M\bar{y}\pm\sqrt{M(1+\frac{M}{N})\bar{y}}.\end{aligned} \tag{318}$$

*Equation* (318) *informs that the expected number of future events is equal to the mean number of observed events per day multiplied with the number of days – That makes sense. The variance of equation* (318) *is proportional to the square root of the observed mean event count, which is to be expected from a Poisson distribution. It means* $\mathbb{E}[x|y,I]\gg\sqrt{Var[x|y,I]}$ *for* $\bar{y}\gg 1$ *and hence that the* *uncertainty approaches a negligible quantity. This is a strong implicit statement which easily can be broken. This can for example happen if the number of events is expected to vary due to a rate parameter that changes every day due to underlying dynamics, e.g. weather conditions.*

advanced poisson assumption: *In the case where the rate parameter of the Poisson distribution is expected to have significant time variance while retaining unimodality, it is more accurate to assume that each rate parameter is unique, but drawn from the same gamma distribution, meaning*

$$\begin{aligned} p(x=i|y,\xi,\zeta,I) &= \int d\{\lambda_x\}d\{\lambda_y\}d\alpha d\beta p(x=i,\{\lambda_x\},\{\lambda_y\},\alpha,\beta|y,\xi,\zeta,I) \\ &= \int d\{\lambda_x\}d\{\lambda_y\}d\alpha d\beta p(x=i,\{\lambda_x\}|\alpha,\beta,y,\xi,\zeta,I)p(\{\lambda_y\},\alpha \end{aligned} \tag{319}$$

*where $\xi,\zeta$ are the parameters of the prior distributions, $d\{\lambda_x\} = d\lambda_{x_1}d\lambda_{x_2}\dots d\lambda_{x_M}$ and $d\{\lambda_y\} = d\lambda_{y_1}d\lambda_{y_2}\dots d\lambda_{y_N}$ and*

$$\begin{aligned} p(x=i,\{\lambda_x\}|\alpha,\beta,y,\xi,\zeta,I) &= p(x=i|\{\lambda_x\},\alpha,\beta,y,I)p(\{\lambda_x\}|\alpha,\beta,I), \\ p(\{\lambda_y\},\alpha,\beta|y,\xi,\zeta,I) &= \frac{p(y|\{\lambda_y\},\alpha,\beta,I)p(\{\lambda_y\}|\alpha,\beta,I)p(\alpha|\xi,I)p(\beta|\zeta}{p(y|\xi,\zeta,I)} \end{aligned} \tag{320}$$

*with*

$$\begin{aligned} p(y|\xi,\zeta,I) &= \int d\{\lambda_y\}d\alpha d\beta p(y|\{\lambda_y\},\alpha,\beta,I)p(\{\lambda_y\}|\alpha,\beta \\ p(y|\{\lambda_y\},\alpha,\beta,I)p(\{\lambda_y\}|\alpha,\beta,I) &= \prod_{j=1}^{N} Poi(y_j|\lambda_{y_j})Ga(\lambda_{y_j}|\alpha,\beta). \end{aligned} \tag{321}$$

*The integrals over gamma can be evaluated individually viz*

$$\begin{aligned} p(y_i|\alpha,\beta,I) &= \int d\lambda_{y_i} Poi(y_j|\lambda_{y_j}) Ga(\lambda_{y_j}|\alpha,\beta) \\ &= \binom{y_j+\alpha-1}{y_j}\left(\frac{\beta}{\beta+1}\right)^{y_j}\left(\frac{1}{\beta+1}\right)^{\alpha} \\ &= NB\left(y_j|\alpha,\frac{\beta}{\beta+1}\right), \end{aligned} \tag{322}$$

*where NB abbreviates the negative binomial distribution. The sum of negative binomial random variables is another negative binomial random variable, such that*

$$\int d\{\lambda_x\} p(x=i,\{\lambda_x\}|\alpha,\beta,y,I) = NB\left(x=i|M\alpha,\frac{\beta}{\beta+1}\right), \tag{323}$$

*meaning equation* (319) *can be written*

$$p(x=i|y,\xi,\zeta,I) = \frac{1}{p(y|\xi,\zeta,I)}\int d\alpha d\beta NB\left(x=i|M\alpha,\frac{\beta}{\beta+1}\right)\prod_{j=1}^{N} NB\left(y_j|\alpha, \right. \tag{324}$$

*Using the principle of maximum, entropy, the distributions for $\alpha$ and $\beta$ can be assigned gamma distributions*

$$\begin{aligned} p(\alpha|\xi,I) &= p(\alpha|a,b,I) \\ &= Ga(\alpha|a,b), \\ p(\beta|\zeta,I) &= p(\beta|c,d,I) \\ &= Ga(\alpha|c,d), \end{aligned} \tag{325}$$

*with $\xi = \{a,b\}$ and $\zeta = \{c,d\}$. With this setup, $\mathbb{E}[x|y,\xi,\zeta,I] \pm \sqrt{Var[x|y,\xi,\zeta,I]}$ can be evaluated numerically using numerical methods like e.g. Hamiltonian Monte Carlo (see appendix A). In this example, the pymc python packge [52] will be used (which utilize Hamiltonian Monte Carlo) with parameters $a = b = c = d = 1$, corresponding to wide (low bias) gamma distributions.*

NORMAL ASSUMPTION: *In case the distribution of x is fairly symmetric around its peak and the summary statistics of equation* (309) *are the quantities of primary interest, a crude normal approximation can be made. For a continuous variable, equation* (310) *become*

$$\begin{aligned} \mathbb{E}[x|y,I] &= \int xp(x|y,I), \\ Var[x|y,I] &= \int x^2 p(x|y,I) - \mathbb{E}[x|y,I]^2. \end{aligned} \tag{326}$$

*Assume x is normally distributed with mean and precision* $\mu \sim N(\mu|mean = \mu_0, variance = \frac{1}{c\tau})$ *and* $\tau \sim Ga(\tau|\alpha,\beta)$ *following a normal and gamma distribution, respectively. In this case*

$$p(x|y,\mu_0,c,\alpha,\beta,I) = \int d\mu d\tau p(x|\mu,\tau,y,I) \frac{p(y|\mu,\tau,I)p(\mu|\tau,\mu_0,c,I)p(\tau|\alpha,\beta,}{p(y|\mu_0,c,\alpha,\beta,I)} \tag{327}$$

*With these assumptions, the likelihood is given by*

$$\begin{aligned} p(y|\mu,\tau,I) &= \left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}} \prod_{j=1}^{N} e^{-\frac{\tau}{2}(\mu-y_j)^2} \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2} \end{aligned} \tag{328}$$

*and the expectation can be written*

$$\begin{aligned} \mathbb{E}[x|y,\mu_0,c,\alpha,\beta,I] &= \frac{M\beta^\alpha}{p(y|I)\Gamma(\alpha)} \int d\mu d\tau \mu \left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2} \sqrt{\frac{c\tau}{2\pi}} e^{-\frac{c\tau}{2}(\mu} \\ &= M\frac{Z_1}{Z_0}, \end{aligned} \tag{329}$$

*with*

$$\begin{aligned} Z_1 &\equiv \int d\tau \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} e^{-\tau\beta} \int d\mu \mu e^{-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2 - \frac{c\tau}{2}(\mu-\mu_0)^2}, \\ Z_0 &\equiv \int d\tau \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} e^{-\tau\beta} \int d\mu e^{-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2 - \frac{c\tau}{2}(\mu-\mu_0)^2}. \end{aligned} \tag{330}$$

*For the identification of $Z_1$ and $Z_0$ it has been used that all constants cancel out due to $p(y|I)$. The exponents involving $\mu$ can be rewritten*

$$\begin{aligned}
-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2-\frac{c\tau}{2}(\mu-\mu_0)^2 &= -\frac{\tau}{2}\sum_{j=1}^{N}(\mu-\bar{y}+\bar{y}-y_j)^2-\frac{c\tau}{2}(\mu-\mu_0)^2\\
&= -\frac{\tau}{2}\sum_{j=1}^{N}(\bar{y}-y_j)^2-\frac{N\tau(\mu-\bar{y})^2}{2}-\tau(\mu-\bar{y}\\
&= -\frac{\tau}{2}\sum_{j=1}^{N}(\bar{y}-y_j)^2-\frac{N\tau(\mu-\bar{y})^2}{2}-\frac{c\tau}{2}(\mu-\\
&= -\frac{\tau}{2}\sum_{j=1}^{N}(\bar{y}-y_j)^2-\frac{\tau}{2}(N+c)\left[\left(\mu-\frac{c\mu_0+}{N+}\right.\right.
\end{aligned} \tag{331}$$

*Only the second term in equation* (331) *depend on $\mu$, so*

$$e^{-\beta\tau}e^{-\frac{\tau}{2}\sum_{j=1}^{N}(\mu-y_j)^2-\frac{c\tau}{2}(\mu-\mu_0)^2}=e^{-\beta_0\tau}e^{-\frac{\tau}{2}(N+c)(\mu-\frac{c\mu_0+N\bar{y}}{N+c})^2}, \tag{332}$$

*with*

$$\beta_0\equiv\beta+\frac{1}{2}\sum_{j=1}^{N}(\bar{y}-y_j)^2+\frac{cN(\bar{y}-\mu_0)^2}{2(c+N)}. \tag{333}$$

*Using equation* (332) *equation* (330) *can be written*

$$\begin{aligned}
Z_1 &= \int d\tau\tau^{\alpha_0-\frac{1}{2}}e^{-\tau\beta_0}\int d\mu\mu e^{-\frac{\tau}{2}(N+c)(\mu-\frac{c\mu_0+N\bar{y}}{N+c})^2}\\
&= \sqrt{\frac{2\pi}{N+c}}\frac{c\mu_0+N\bar{y}}{N+c}\beta_0^{-\alpha_0}\Gamma(\alpha_0),\\
Z_0 &= \int d\tau\tau^{\alpha_0-\frac{1}{2}}e^{-\tau\beta_0}\int d\mu e^{-\frac{\tau}{2}(N+c)(\mu-\frac{c\mu_0+N\bar{y}}{N+c})^2}\\
&= \sqrt{\frac{2\pi}{N+c}}\beta_0^{-\alpha_0}\Gamma(\alpha_0)
\end{aligned} \tag{334}$$

*where* $\alpha_0 \equiv \alpha + \frac{N}{2}$. *Combining equation* (329) *with* (334) *yields*

$$\mathbb{E}[x|y,\mu_0,c,\alpha,\beta,I] = M\frac{c\mu_0+N\bar{y}}{c+N} \tag{335}$$

*For the variance*

$$\begin{aligned}\mathbb{E}[x^2|y,\mu_0,c,\alpha,\beta,I] &= \int d\mu d\tau dx x^2 p(x|\mu,\tau,y,I)\frac{p(y|\mu,\tau,I)p(\mu|\tau,\mu_0,c,I)p(\tau}{p(y|\mu_0,c,\alpha,\beta,I)}\\ &= \frac{M}{p(y|\mu_0,c,\alpha,\beta,I)}\int d\mu d\tau (M\mu^2+\tau^{-1})p(y|\mu,\tau,I)p(\mu|\\ &= M\frac{MZ_2+Z_3}{Z_0}\end{aligned} \tag{336}$$

*where*

$$\begin{aligned}Z_2 &= \int d\tau \tau^{\alpha_0-\frac{1}{2}}e^{-\tau\beta_0}\int d\mu \mu^2 e^{-\frac{\tau}{2}(N+c)(\mu-\frac{c\mu_0+N\bar{y}}{N+c})^2}\\ &= \sqrt{\frac{2\pi}{(N+c)}}\left[\frac{1}{N+c}\beta_0^{1-\alpha_0}\Gamma(\alpha_0-1)+\left(\frac{c\mu_0+N\bar{y}}{N+c}\right)^2\beta_0^{-\alpha_0}\Gamma(\alpha_0)\right],\\ Z_3 &= \int d\tau \tau^{\alpha_0-\frac{3}{2}}e^{-\tau\beta_0}\int d\mu e^{-\frac{\tau}{2}(N+c)(\mu-\frac{c\mu_0+N\bar{y}}{N+c})^2}\\ &= \sqrt{\frac{2\pi}{(N+c)}}\beta_0^{1-\alpha_0}\Gamma(\alpha_0-1)\end{aligned} \tag{337}$$

*Combining equations* (310), (336) *and* (337)

$$Var[x|y,\mu_0,c,\alpha,\beta,I] = M\left(\frac{M}{N+c}+1\right)\frac{\beta_0}{\alpha_0-1} \tag{338}$$

*In the limit of* $\mu_0, c, \alpha, \beta \to 0$ *then*

$$\begin{aligned}\mathbb{E}[x|y,I] \pm \sqrt{Var[x|y,I]} &= \lim_{\mu_0,c,\alpha,\beta\to 0}\left(\mathbb{E}[x|y,\mu_0,c,\alpha,\beta,I] \pm \sqrt{Var[x|y,\mu_0,c,\alpha,}\right.\\ &= M\bar{y} \pm \sqrt{M\frac{N+M}{N}\frac{N-1}{N-2}\delta y}\end{aligned}$$

(339)

*where*

$$\delta y \equiv \sqrt{\frac{\sum_{j=1}^{N}(\bar{y}-y_j)^2}{N-1}} \tag{340}$$

*is the sample standard deviation. From equation* (339)*, it is clear that the result is close to the sample mean with uncertainty close to the sample standard deviation.*

**NUMERICAL EXAMPLE AND COMPARISON OF APPROXIMATIONS:** *In order to illustrate the difference between the approximations considered in this example, consider a numerical example where data is drawn from a Poisson distribution with a rate parameter that is drawn from a gamma distribution with parameters* $\alpha = 10$ *and* $\beta = 0.01$. $300$ *counts are drawn, yielding the data shown in figure 13. Given the data shown in figure 13, the three approximations return the posterior predicitve distributions (*$M = 1$*) shown in figure 14 with expectations and standard deviations shown in table 1. From figure 14 it is clear that the approximations yield significantly different distributions. From table 1 it is clear that the expectation values are highly similar, whereas the standard deviations differ significantly. The distributions underlying each approximation (Poisson, negative binomial and normal) are unimodal, and as such it is expected that they will estimate the expectation value of unimodal data accurately. The Poisson distribution has a standard deviation that is determined by the expectation value (see equation* (318)*) and as such it cannot accurately describe the variance in data where the expectation and variance is decoupled. In the advanced Poisson approximation, the mean and variance are decoupled and hence an accurate description of data is seen. Relative to the advanced Poisson approximation, the normal approximation suffer from three shortcommings i) it assumes the discrete events are continuous, ii) it assumes the distribution is symmetric and iii) it allows the counts to be negative (see the* *left tail of figure 14 (bottom)). In relation to the expectation and standard deviation neither of the shortcommings are highly significant, however, they may be if different questions are asked of the posterior predictive.*

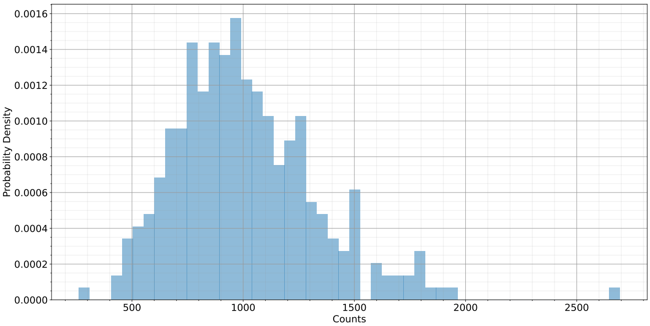

Figure 13: A probability density histogram for the sample of 300 data points considered in the numerical example. The data are event counts drawn from (300) Poisson distributions with (300) rate parameters drawn from a gamma distribution with $\alpha = 10$ and $\beta = 0.01$.

| Approximation | $\mathbb{E}[x\|y, I]$ | $\sqrt{\text{Var}[x\|y, I]}$ |
|---|---|---|
| Simple Poisson theoretical | 1008.9 | 31.8 |
| Simple Poisson pymc | 1008.5 | 31.8 |
| Advanced Poisson pymc | 1008.7 | 324.1 |
| Normal theoretical | 1008.9 | 321.3 |
| Normal pymc | 1007.5 | 322.2 |

Table 1: Equation (309) computed by each of the three approximations considered here.

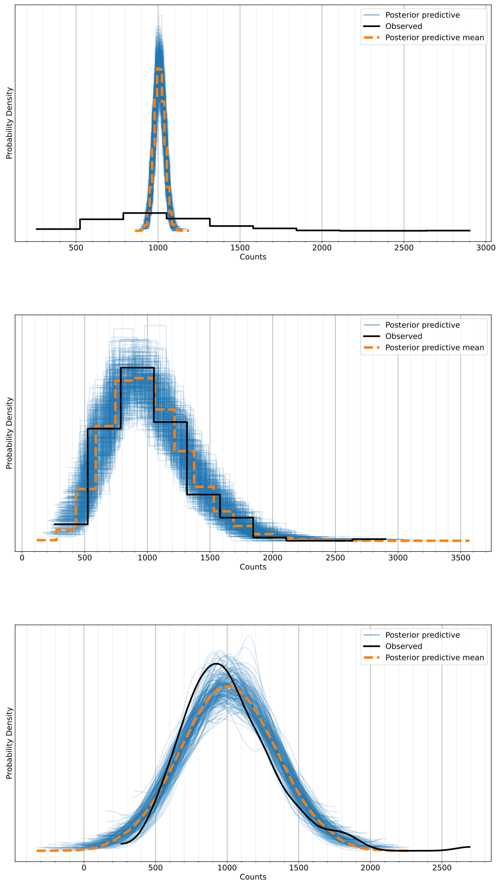

Figure 14: The posterior predictive distributions for the numerical example for each approximation. (top) is the simple Poisson approximation, (middle) is the advanced Poisson approximation and (bottom) is the normal approximation.

# CHAPTER 11

# Making Inference About the Model of Nature

In some instances, the robot is interested in inference related to the model of Nature. The observation $X = x$ by definition does not have an associated known action of Nature and thus by axiom 6 is disregarded in this context. From equation (274)

$$U^* = \arg\min_U \mathbb{E}[C(U,S)|D,I], \tag{341}$$

where $S = s$ is interpreted as an action related to the model of Nature, e.g. Nature picking a given systematic that generates data.

## 11.1 selecting the robot's model

Suppose the Robot must choose between two competing models, aiming to select the one that best represents Nature's true model. The two competing models could e.g. be two different functions $f$ in regression or two different probability distribution assignments. In this case the Robot has actions $u_1$ and $u_2$ representing picking either model and Nature has two actions $s_1$ and $s_2$ which represent which model that in truth fit Nature's true model best. From equation (274)

$$\begin{aligned}\mathbb{E}[C(u_1,S)|D,I] &= \sum_{s=s_1,s_2} C(u_1,s)p(S=s|D,I),\\ \mathbb{E}[C(u_2,S)|D,I] &= \sum_{s=s_1,s_2} C(u_2,s)p(S=s|D,I),\end{aligned} \tag{342}$$

where in this case $u_i = s_i \quad \forall(u_i,s_i) \in \mathbb{U} \times \mathbb{S}$ but the notational distinction is kept to avoid confusion. Since there is no input $X = x$ in this case, the decision rule $U$ is fixed (i.e. it does not depend on $x$). $U = u_1$ is picked iff $\mathbb{E}[C(U = u_1,S)|D,I] < \mathbb{E}[C(U = u_2,S)|D,I]$, meaning

$$\frac{p(s_1|D,I)}{p(s_2|D,I)} > \frac{C(u_1,s_2) - C(u_2,s_2)}{C(u_2,s_1) - C(u_1,s_1)}. \tag{343}$$

The ratio $\frac{p(s_1|D,I)}{p(s_2|D,I)}$ is referred to as the posterior ratio. Using Bayes theorem it can be re-written viz

$$\begin{aligned}\text{posterior ratio} &= \frac{p(s_1|D,I)}{p(s_2|D,I)}\\ &= \frac{p(D_s|s_1,D_x,I)p(s_1|I)}{p(D_s|s_2,D_x,I)p(s_2|I)},\end{aligned} \tag{344}$$

where for the second equality it has been used that the normalization $p(D|I)$ cancels out between the denominator and nominator and axiom 6 has been employed. Given there is no a priori bias towards any model, $p(s_1|I) = p(s_2|I)$

$$\text{posterior ratio} = \frac{p(D_s|s_1,D_x,I)}{p(D_s|s_2,D_x,I)}. \tag{345}$$

$p(D_s|s_1,D_x,I)$ and $p(D_s|s_2,D_x,I)$ can then be expanded via marginalization, the chain rule and Bayes theorem until they can be evaluated either analytically or numerically. Equation (345) is referred to as Bayes factor and as a rule of thumb

**Definition 55** (Bayes Factor Interpretation Rule of Thumb)**.** *If the probability of either of two models being the model of Nature is more than 3 times likely than the other, the likelier model is accepted. Otherwise the result does not significantly favor either model.*

## 11.2 PARAMETER ESTIMATION

Let $w_j \in \mathbb{W}$ represent the $j'$th parameter with the associated random variable $W_j$. In case of parameter estimation, the action of Nature is identified with the parameter of interest from the model of Nature's and the Robot's action with the act of estimating the parameters value, meaning

$$U^* = \arg\min_U \mathbb{E}[C(U, W_j)|D, I], \tag{346}$$

with

$$\mathbb{E}[C(U, W_j)|D, I] = \int dsC(U, w_j)p(w_j|D, I). \tag{347}$$

At this point, the Robot can select a cost function like in section 10 and proceed by expanding $p(w_j|D, I)$ similarly to equation (170). Picking the quadratic cost yields

$$U^* = \mathbb{E}[w_j|D, I] \tag{348}$$

$p(w_j|D, I)$ in (348) can be expanded as shown in equation (170).

**Example 11.1.**

*Consider the scenario where two sets of costumers are subjected to two different products, A and B. After exposure to the product, the costumer will be asked whether or not they are satisfied and they will be able to answer "yes" or "no" to this. Denote the probability of a costumer liking product $A/B$ by $w_A/w_B$, respectively. In this context, the probabilities $w_A/w_B$ are parameters of Natures model (similar to how the probability is a parameters for a binomial distribution). What will be of interest is the integral of the joint probability distribution where $w_B > w_A$, meaning*

$$p(w_B > w_A|D, I) = \int_0^1 \int_{w_A}^1 p(w_A, w_B|D, I) dw_A dw_B. \quad (349)$$

*Assuming the costumer sets are independent*

$$\begin{aligned} p(w_A, w_B|D, I) &= p(w_B|w_A, D, I)p(w_A|D, I) \\ &= p(w_B|D_A, I)p(w_A|D_A, I), \end{aligned} \quad (350)$$

*with*

$$p(w_i|D_i, I) = \frac{p(D_i|w_i, I)p(w_i|I)}{p(D_i|I)}. \quad (351)$$

*Assuming a beta prior and a binomial likelihood yields (since the binomial and beta distributions are conjugate)*

$$p(w_i|D_i, I) = \frac{w_i^{\alpha_i - 1}(1 - w_i)^{\beta_i - 1}}{B(\alpha_i, \beta_i)}, \quad (352)$$

*where $\alpha_i \equiv \alpha + s_i$, $\beta_i \equiv \beta + f_i$ and $s_i/f_i$ denotes the successes/failure, respectively, registered in the two sets of costumers. Evaluating equation* (349) *yields*

$$p(w_B > w_A|D, I) = \sum_{j=0}^{\alpha_B - 1} \frac{B(\alpha_A + j, \beta_A + \beta_B)}{(\beta_B + j)B(1 + j, \beta_B)B(\alpha_A, \beta_A)}. \quad (353)$$

**Example 11.2.**

Consider a uniform distribution centered on 0 with width $2a$. The density distribution is given by $p(x|a) = \frac{\mathbb{I}(x \in [-a,a])}{2a}$.

1. Given a dataset $x_1, x_2, \ldots x_n$ what is the MLE estimate of $a$?

$$\begin{aligned} p(D|a,I) &= \prod_{j=1}^{n} (\frac{\mathbb{I}(x_j \in [-a,a])}{2a}) \\ &= \frac{1}{(2a)^n} \textit{ for } x_j \in [-a,a] \end{aligned} \tag{354}$$

*The likelihood cannot be optimized using the derivative approach since* $\frac{d \ln(p(D|a,I))}{da}\Big|_{a=\hat{a}_{MLE}} = 0$ *does not yield a useable result. Instead note that* $(2a)^{-n}$ *is a monotonic decreasing function of a, meaning the likelihood will be largest for the smallest value of a whilst obeying* $x_j \leq a$. *This means* $\hat{a}_{MLE} = \max(|x_1|, |x_2|, \ldots |x_n|)$, *since this will yield the smallest value of a that is still larger or equal to all the data values.*

2. What probability would the model assign to a new datapoint $x_{n+1}$ using $\hat{a}_{\text{MLE}}$?

$$p(x_{n+1}|D,I)|_{MLE} = \frac{1}{2\max(|x_1|, |x_2|, \ldots |x_n|)} \tag{355}$$

*for* $-\max(|x_1|, |x_2|, \ldots |x_n|) \leq x_{n+1} \leq \max(|x_1|, |x_2|, \ldots |x_n|)$.

3. Do you see any problems with the above approach? Briefly suggest a better approach.

*The model place zero probability mass outside the training data. You need to use prior information.*

**Example 11.3.**

Suppose we have two sensors with unknown variances $\nu_1$ and $\nu_2$ ($\nu_1 \neq \nu_2$), but unknown (and the same) mean, $\mu$. Suppose we observe $n_1$ observation from $y_i^{(1)} \sim N(y_i^{(1)}|\mu, \nu_1)$ from the first sensor and $y_i^{(2)} \sim N(y_i^{(2)}|\mu, \nu_2)$ from the second sensor. Let $D$ represent all the data from both sensors. What is the posterior $p(\mu|D, \nu_1, \nu_2, I)$, assuming a non-informative prior for $\mu$?

$$p(\mu|D, \nu_1, \nu_2, I) = \frac{p(D|\mu, \nu_1, \nu_2, I)p(\mu|\nu_1, \nu_2, I)}{p(D|\nu_1, \nu_2, I)} \tag{356}$$

*with $p(\mu|\nu_1, \nu_2, I) = Unif(a, b)$ since it is supposed to be an uninformative prior. The likelihood*

$$\begin{aligned} p(D|\mu, \nu_1, \nu_2, I) &= \prod_{i=1}^{N^{(1)}} \frac{1}{\sqrt{2\pi\nu_1}} e^{-\frac{1}{2\nu_1}(y_i^{(1)}-\mu)^2} \prod_{j=1}^{N^{(2)}} \frac{1}{\sqrt{2\pi\nu_2}} e^{-\frac{1}{2\nu_2}(y_j^{(2)}-\mu)^2} \\ &= (2\pi\nu_1)^{-\frac{N^{(1)}}{2}} (2\pi\nu_2)^{-\frac{N^{(2)}}{2}} e^{-\frac{1}{2\nu_1}\sum_{i=1}^{N^{(1)}}(y_i^{(1)}-\mu)^2 - \frac{1}{2\nu_2}\sum_{j=1}^{N^{(2)}}(y_j^{(2)}-\mu)^2} \\ &\propto e^{-\frac{1}{2\tilde{\nu}}(\mu-\tilde{\mu})^2}, \end{aligned} \tag{357}$$

*where $\tilde{\nu}$ can be found by considering the part of the exponent for $-\frac{\mu^2}{2\tilde{\nu}}$ and then identifying $\tilde{\nu}$, yielding*

$$\tilde{\nu} = \frac{1}{\frac{N^{(1)}}{\nu_1} + \frac{N^{(2)}}{\nu_2}}. \tag{358}$$

*Similarly by considering the part of the exponent for the double product*

$$\tilde{\mu} = \tilde{\nu}\left(\frac{N^{(1)}\bar{y}^{(1)}}{\nu_1} + \frac{N^{(2)}\bar{y}^{(2)}}{\nu_2}\right). \tag{359}$$

# Part IV

# REFLECTION

# CHAPTER 12

## Reflections on Statistical Paradigm

Bayesian statistics has mathematical beauty in the way of physics models. It is more appealing – you can compare the situation to general relativity and Newtonian physics. The former is conceptually simpler and objectively better, even in the Newtonian limit. However, it is mathematically heavy and therefore people use Newtonian physics in this limit. Newtonian physics can be boiled down to a tool much better than GR. The same goes for Frequentist statistics. I believe this is why it is used more, as well as the computational aspect. If technology allows Bayesian statistics to better boil down to a tool, then I believe Frequentist statistics will be a thing of history.

# Part V

# APPENDIX

# APPENDIX A

# Hamiltonian Monte Carlo

This appendix is taken from Petersen [53]. The Hamiltonian Monte Carlo Algorithm (HMC algorithm) is a Markov Chain Monte Carlo (MCMC) algorithm used to evaluate integrals on the form

$$\begin{aligned} \mathbb{E}[f] &= \int f(\theta)g(\theta)d\theta \\ &\approx \frac{1}{N}\sum_{j\in g} f(\theta_j), \end{aligned} \tag{360}$$

with $f$ being a generic function and $N$ denoting the number of samples from the posterior distribution, $g$. The sample $\{j\}$ from $g$ can be generated via a MCMC algorithm that has $g$ as a stationary distribution. The Markov chain is defined by an initial distribution for the initial state of the chain, $\theta$, and a set of transition probabilities, $p(\theta'|\theta)$, determining the sequential evolution of the chain. A distribution of points in the Markov Chain are said to comprise a stationary distribution if they are drawn from the same distribution and that this distribution persist once established. Hence, if $g$ is the a stationary distribution of the Markov Chain defined by the initial point $\theta$ and the transition probability $p(\theta'|\theta)$, then [36]

$$g(\theta') = \int p(\theta'|\theta)g(\theta)d\theta. \tag{361}$$

Equation (361) is implied by the stronger condition of detailed balance, defined viz

$$p(\theta'|\theta)g(\theta) = p(\theta|\theta')g(\theta'). \tag{362}$$

A Markov chain is ergodic if it has a unique stationary distribution, called the equilibrium distribution, to which it converge from any initial state. $\{i\}$ can be taken as a sequential subset (discarding the part of the chain before the equilibrium distribution) of a Markov chain that has $g(\theta)$ as its equilibrium distribution.

The simplest MCMC algorithm is perhaps the Metropolis-Hastings (MH) algorithm [54, 55]. The MH algorithm works by randomly initiating all coefficients for the distribution wanting to be sampled. Then, a loop runs a subjective number of times in which one coefficient at a time is perturbed by a symmetric proposal distribution. A common choice of proposal distribution is the normal distribution with the coefficient value as the mean and a subjectively chosen variance. If $g(\theta') \geq g(\theta)$ the perturbation of the coefficient is accepted, otherwise the perturbation is accepted with probability $\frac{g(\theta')}{g(\theta)}$.

The greatest weaknesses of the MH algorithm is i) a slow approach to the equilibrium distribution, ii) relatively high correlation between samples from the equilibrium distribution and iii) a relatively high rejection rate of states. ii) can be rectified by only accepting every $n$'th accepted state, with $n$ being some subjective number. For $n \to \infty$ the correlation naturally disappears, so there is a trade off between efficiency and correlation. Hence, in the end the weaknesses of the MH algorithm can be boiled down to inefficiency. This weakness is remedied by the HCM algorithm [35] in which Hamiltonian dynamics are used to generate proposed states in the Markov chain and thus guide the journey in parameter space. Hamiltonian dynamics are useful for proposing states because [37] 1) the dynamics are reversible, implying that detailed balance is fulfilled and so there exist a stationary distribution, 2) the Hamiltonian ($H$) is conserved during the dynamics if there is no explicit time dependence in the Hamiltonian ($\frac{dH}{dt} = \frac{\partial H}{\partial t}$), resulting in all proposed states being accepted in the case the dynamics are exact and 3) Hamiltonian dynamics preserve the volume in phase space ($q_i, p_i$-space), which means that the Jacobian is unity (relevant for Metropolis updates that succeeds the Hamiltonian dynamics in the algorithm). By making sure the algorithm travel (in parameter space) a longer distance between proposed states, the proposed states can be ensured to have very low correlation, hence alleviating issues 1) and 2) of the MH algorithm. The price to pay for using the HMC algorithm relative to the MH algorithm is a) the HMC algorithm is gradient based meaning it requires the Hamiltonian to be continuous and b) the computation time can be long depending on the distribution being sampled (e.g. some recurrent ANNs are computationally heavy due to extensive gradient calculations).

As previously stated, the HMC algorithm works by drawing a physical analogy and using Hamiltonian dynamics to generate proposed states and thus guide the journey in parameter space. The analogy consists in viewing $g$ as the canonical probability distribution describing the probability of a given configuration of parameters. In doing so, $g$ is related to the Hamiltonian, $H$, viz

$$g = e^{\frac{F-H}{k_B T}} \Rightarrow H = F - k_B T \ln[g], \tag{363}$$

where $F = -k_B T ln[Z]$ denotes Helmholtz free energy of the (fictitious in this case) physical system and $Z$ is the partition function. $\ln[g(\theta)]$ contain the position (by analogy) variables of the Hamiltonian and so $Z$ must contain the momentum variables. Almost exclusively [56] $Z \sim \mathcal{N}(0, \sqrt{m_i})$ is taken yielding the Hamiltonian

$$H = -k_B T \left[ \ln[g] - \sum_i \frac{p_i^2}{2m_i} \right] + const, \tag{364}$$

where $i$ run over the number of variables and "const" is an additive constant (up to which the Hamiltonian is always defined). $T = k_b^{-1}$ is most often taken [37], however, the temperature can be used to manipulate the range of states which can be accepted e.g. via simulated annealing [57]. Here $T = k_b^{-1}$ will be adopted in accordance with [36, 37] and as such

$$H = \sum_i \frac{p_i^2}{2m_i} - \ln[g]. \tag{365}$$

The dynamics in parameter space are determined by Hamiltons equations

$$\dot{\theta}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial \theta_i}, \tag{366}$$

with $\theta_i$ denoting the different variables (coefficients). In order to implement Hamiltons equations, they are discretized via the leap frog method [36, 37] viz

$$\begin{aligned} p_i\left(t+\frac{\epsilon}{2}\right) &= p_i(t) - \frac{\epsilon}{2}\frac{\partial H(\theta_i(t), p_i(t))}{\partial \theta_i}, \\ \theta_i(t+\epsilon) &= \theta_i(t) + \frac{\epsilon}{m_i} p_i\left(t+\frac{\epsilon}{2}\right), \\ p_i(t+\epsilon) &= p_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial H(\theta_i(t+\frac{\epsilon}{2}), p_i(t+\frac{\epsilon}{2}))}{\partial \theta_i}, \end{aligned} \tag{367}$$

with $\epsilon$ being an infinitesimal parameter. In the algorithm the initial state is defined by a random initialization of coordinates and momenta, yielding $H_{initial}$. Subsequently Hamiltonian dynamics are simulated a subjective ($L$ loops) amount of time resulting in a final state, $H_{final}$, the coordinates of which take the role of proposal state. The loop that performs $L$ steps of $\epsilon$ in time is here referred to as the dive. During the dive, the Hamiltonian remains constant, so ideally $H_{initial} = H_{final}$, however, imperfections in the discretization procedure of the dynamics can result in deviations from this equality (for larger values of $\epsilon$, as will be discussed further later on). For this reason, the proposed state is accepted as the next state in the Markov chain with probability

$$\mathbb{P}(\text{transition}) = \min\left[1, e^{H_{initial}-H_{final}}\right]. \tag{368}$$

Whether or not the proposed state is accepted, a new proposed state is next generated via Hamiltonian dynamics and so the loop goes on for a subjective amount of time. Most often, the HMC algorithm will be ergodic, meaning it will converge to its unique stationary distribution from any given initialization (i.e. the algorithm will not be trapped in some subspace of parameter space), however, this may not be so for a periodic Hamiltonian if $L\epsilon$ equal the periodicity. This potential problem can however be avoided by randomly choosing $L$ and $\epsilon$ from small intervals for each iteration. The intervals are in the end subjective, however, with some constraints and rules of thumb; the leap frog method has an error of $\mathcal{O}(\epsilon^2)$ [36] and so the error can be controlled by ensuring that $\epsilon \ll 1$. A too small value of $\epsilon$ will waste computation time as a correspondingly larger number of iterations in the dive ($L$) must be used to obtain a large enough trajectory length $L\epsilon$. If the trajectory length is too short the parameter space will be slowly explored by a random walk instead of the otherwise approximately independent sampling (the advantage of non-random walks in HMC is a more uncorrelated Markov chain and better sampling of the parameter space). A rule of thumb for the choice of $\epsilon$ can be derived from a one dimensional Gaussian Hamiltonian

$$H = \frac{q^2}{2\sigma^2} + \frac{p^2}{2}. \tag{369}$$

The leap frog step for this system is a linear map from $t \to t+\epsilon$. The mapping can be written

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \begin{bmatrix} 1-\frac{\epsilon^2}{2\sigma^2} & \epsilon \\ \epsilon(\frac{1}{4}\epsilon^2\sigma^{-4} - \sigma^{-2}) & 1-\frac{1}{2}\epsilon^2\sigma^{-2} \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \tag{370}$$

The eigenvalues of the coefficient matrix represent the powers of the exponentials that are the solutions to the differential equation. They are given by

$$\text{Eigenvalues} = 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \pm \epsilon\sigma^{-1}\sqrt{\frac{1}{4}\epsilon^2\sigma^{-2} - 1}. \quad (371)$$

In order for the solutions to be bounded, the eigenvalues must be imaginary, meaning that

$$\epsilon < 2\sigma. \quad (372)$$

In higher dimensions a rule of thumb is to take $\epsilon \lesssim 2\sigma_x$, where $\sigma_x$ is the standard deviation in the most constrained direction, i.e. the square root of the smallest eigenvalue of the covariance matrix. In general [56] a stable solution with $\frac{1}{2}p^T\Sigma^{-1}p$ as the kinetic term in the Hamiltonian require

$$\epsilon_i < 2\lambda_i^{-\frac{1}{2}}, \quad (373)$$

for each eigenvalue $\lambda_i$ of the matrix

$$M_{ij} = (\Sigma^{-1})_{ij}\frac{\partial^2 H}{\partial q_i \partial q_j}, \quad (374)$$

which means that in the case of $\Sigma^{-1} = diag(m_i^{-1})$;

$$\epsilon_i < 2\sqrt{\frac{m_i}{\frac{\partial^2 H}{\partial q_i^2}}}. \quad (375)$$

Setting $\epsilon$ according to equation (373) can however introduce issues for hierarchical models (models including hyper parameters) since the reversibility property of Hamiltonian dynamics is broken if $\epsilon$ depend on any parameters. This issue can be alleviated by using the MH algorithm on a subgroup of parameters [36, 37] (which are then allowed in the expression for $\epsilon$) that is to be included in $\epsilon$. However, unless the MH algorithm is used for all parameters, some degree of approximation is required.

**Algorithm 4** Hamiltonian Monte Carlo Algorithm in pseudo code

```
Save: $q$ and $V(q)$, with $q$ randomly initialized
for $i \leftarrow 1$ to $N$ do
    $p \leftarrow$ Sample from standard normal distribution
    $H_{\text{old}} \leftarrow H(q,p)$
    $p \leftarrow p - \frac{\epsilon}{2}\frac{\partial H(q,p)}{\partial q}$
    $L \leftarrow$ Random integer between $L_{\text{lower}}$ and $L_{\text{upper}}$
    for $j \leftarrow 1$ to $L$ do
        $q \leftarrow q + \epsilon\frac{p}{\text{mass}}$
        if $j \neq L$ then
            $p \leftarrow p - \epsilon\frac{\partial H(q,p)}{\partial q}$
        end if
    end for
    $p \leftarrow p - \frac{\epsilon}{2}\frac{\partial H(q,p)}{\partial q}$
    $H_{\text{new}} \leftarrow H(q,p)$
    $u \leftarrow$ Sample from uniform distribution
    if $u < \min(1, e^{-(H_{\text{new}} - H_{\text{old}})})$ then
        $H_{\text{old}} \leftarrow H_{\text{new}}$
        Save: $q$ and $V(q)$
    end if
end for
```

# APPENDIX B

# Nested Sampling

A major challenge in estimating the evidence via conventional Monte Carlo Methods is that generally the prior is a very broad and regular distribution whereas the likelihood is a very narrow and irregular distribution. This poses a challenge when the evidence is estimated conventionally, i.e. as the mean of the likelihood evaluated at points in parameter space corresponding to samples from the prior distribution. For a reasonable number of samples, the conventional procedure has a relatively high likelihood of relatively poor sampling in regions near the peaks in the likelihood distribution. This means a conventional estimate of the evidence via Monte Carlo Methods has a high variance. Nested Sampling [58] (NS) address this challenge by accounting for the likelihood distribution when sampling the prior distribution. Consider the integral

$$Z = \int L(\theta)\pi(\theta)d\theta, \tag{376}$$

with $L$ being the likelihood distribution and $\pi$ the prior distribution. Conventional Monte Carlo methods approximate this integral via importance sampling, meaning

$$\begin{aligned} Z &= \mathbb{E}_\pi[L] \\ &\approx \frac{1}{N}\sum_{i\in\pi} L(\theta_i)' \end{aligned} \tag{377}$$

where the second equality become exact for $N \to \infty$. NS project the integral down into one dimension viz[1]

$$\begin{aligned} Z &= \int_0^1 L(\xi) d\xi \\ &\approx \sum_i L(\xi_i) \Delta\xi_i' \end{aligned} \tag{378}$$

where

$$\xi(\lambda) = \int_{L>\lambda} \pi(\theta) d\theta, \tag{379}$$

is the proportion of the prior with likelihood greater than $\lambda$ and $\Delta\xi_i \equiv \xi_{i-1} - \xi_i$. Due to the constraint $L > \lambda$ on the integral bound of $\xi$, $L(\xi)$ is a decreasing function of $\xi$, meaning $L(\xi_1) > L(\xi_2)$ if $\xi_1 < \xi_2$. The sum in equation (378) can then be evaluated by generating a sequence

$$\{\{L(\xi_m), \xi_m\}, \{L(\xi_{m-1}), \xi_{m-1}\}, \dots \{L(\xi_1), \xi_1\}\}, \tag{380}$$

with $\xi_1 < \xi_2 < \cdots < \xi_m$. The sorting operation eliminate coordinate dependent complications of geometry, topology and dimensionality [59]. A sequence upholding equation (380) can be generated as follows; consider $n$ random draws from $g$ with corresponding values of $L$ and $\xi$. Let $L(\xi^*)$ denote the minimum value of $L$ in the sample with $\xi^*$ the corresponding value of $\xi$ in the sample. $\{L(\xi^*), \xi^*\}$ is replaced by another set which is sampled from $g$ with the constraint that $\xi_{new} < \xi^*$ and stored in a list of discarded states. Continuing this sequence again and again will fill the list of discarded states that uphold equation (380). In practice $L(\xi)$ is not readily available, so instead $L$ can be generated from values of $\theta$. The value of $\xi_k$ can be determined by using that [58]

$$\xi_k = \xi_0 \prod_{i=1}^{k} t_i, \tag{381}$$

[1] Attempting a higher accuracy via better numerical approximations of the integral is mute since the uncertainty in $\xi$ dominate the approximation [58].

with $t_i = \frac{\xi_k}{\xi_{k-1}}$, called the shrinkage ratio. The shrinkage ratio follow a beta distribution

$$p(t) = nt^{n-1}, \tag{382}$$

with $n$ being the number of initialy samples from $g$ (the number of live points), such that

$$\begin{aligned} \langle \ln(t) \rangle &= \mathbb{E}[\ln(t)] \pm \sqrt{V[\ln(t)]} \\ &= \int_0^1 nt^{n-1} \ln(t) dt \pm I_2\,, \\ &= \frac{1}{n}(-1 \pm 1) \end{aligned} \tag{383}$$

with

$$I_2 = \sqrt{\int_0^1 nt^{n-1} \ln(t)^2 dt - \left( \int_0^1 nt^{n-1} \ln(t) dt \right)^2}\,. \tag{384}$$

Using $\ln(\xi_k) = \sum_{i=1}^k \ln(t_i)$ and taking $t_i$ to be i.i.d. yield

$$\begin{aligned} \langle \ln(\xi_k) \rangle &= k\mathbb{E}[\ln(t)] \pm \sqrt{kV[\ln(t)]} \\ &= \frac{1}{n}(-k \pm \sqrt{k}) \end{aligned}\,. \tag{385}$$

Ignoring uncertainty $\xi_k$ can be approximated by the mean viz

$$\xi_k \approx e^{-\frac{k}{n}}, \tag{386}$$

meaning

$$\Delta\xi_i \approx e^{-\frac{i}{n}} \left( e^{\frac{1}{n}} - 1 \right). \tag{387}$$

A heuristic measure for terminating the collection of samples is to require that the maximum likelihood collected make up only a small fraction, $B$, of the evidence, meaning

$$\max(\{L\})\xi_j < BZ, \tag{388}$$

for iteration $j$. Another approach to terminating the collection of samples is to use that most of the area in the $L\xi$-plane is usually found in the region [58, 59] $\xi \sim e^{-\mathcal{H}} \sim e^{-\frac{i}{n}}$, meaning the collection of samples can be terminated when

$$i \gg n\mathcal{H}, \tag{389}$$

with $\mathcal{H}$ being the information [58]

$$\begin{aligned} \mathcal{H} &= \int \frac{L(\xi)}{Z} \ln\left(\frac{L(\xi)}{Z}\right) d\xi \\ &\approx \sum_i \frac{L(\xi_i)}{Z} \ln\left(\frac{L(\xi_i)}{Z}\right) \Delta\xi_i. \end{aligned} \tag{390}$$

Temrinating at $i \sim n\mathcal{H}$ yield (equation (385)) an uncertainty $\delta(\langle \ln(\xi_i) \rangle) = \sqrt{\frac{\mathcal{H}}{n}}$ meaning

$$\ln(Z) \approx \ln\left(\sum_i L(\xi_i)\Delta\xi_i\right) \pm \sqrt{\frac{\mathcal{H}}{n}}. \tag{391}$$

The NS algorithm with equation (389) as termination criterion is shown in algorithm 5. $A$ and $B$ are parameters of the algorithm. The "Remainder" in the second to last line in algorithm 5 fills in the missing band $0 < \xi < e^{-\frac{k+1}{n}}$ with the average value of the remaining values of $L$. Due to the chosen stopping criterion, the "Remainder" will be construction be small.

**Algorithm 5** Nested Sampling Algorithm in pseudo code

```
**Import:** $S = n$ samples $\theta_1, \theta_2, \ldots, \theta_n$ from the prior distribution with $L$ being the corresponding likelihoods
**Initialize:** $k \leftarrow 0$, $a \leftarrow 0$, $B \leftarrow 1$, $Z \leftarrow$ Empty list
**while** $f > B$ **do**
    Let $L^* \equiv \min(L)$ and $S^* \widehat{=} L^*$
    $S2 \leftarrow S \setminus S^*$ and $L2 \leftarrow L \setminus L^*$
    Define $\Delta\xi_k = e^{\frac{k+1}{n}}(e^{\frac{1}{n}} - 1)$
    Store $L^* \Delta\xi_k$ in $Z$
    $S_{new}, L_{new} \leftarrow \text{proposer}(\text{random}(S2), L^*)$
    $S \leftarrow S2 \cup S_{new}$ and $L \leftarrow L2 \cup L_{new}$
    $f \leftarrow \frac{\max(L) e^{-\frac{k+1}{n}}}{\sum_{s=0}^{k} Z_s}$
    **if** $a == A$ **then**
        Display status, e.g. $f$, $n\mathcal{H} - k$, $k$, $\sum_{s=0}^{k} Z_s$, ...
        $a \leftarrow 0$
    **end if**
    $k \leftarrow k + 1$
    $a \leftarrow a + 1$
**end while**
Remainder $\leftarrow \frac{1}{n} \sum_i L_i e^{-\frac{k+1}{n}}$
$Z \approx \sum_{s=0}^{k} Z_s +$ Remainder
```

# Bibliography

[1] D. S. Sivia and J. Skilling. *Data Analysis - A Bayesian Tutorial*. 2nd. Oxford Science Publications. Oxford University Press, 2006.

[2] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: `http://probml.github.io/book2`.

[3] Steven M. Lavalle. *Planning Algorithms*. Cambridge University Press, 2006. ISBN: 0521862051.

[4] S.H. Chan. *Introduction to Probability for Data Science*. Michigan Publishing, 2021. ISBN: 9781607857464. URL: `https://books.google.dk/books?id=GR2jzgEACAAJ`.

[5] Andrey Kolmogorov. *Foundations of the Theory of Probability*. Providence, RI, USA: Chelsea Publishing Company, 1950.

[6] Marco Taboga. *Expected value and the Lebesgue integral*. Online appendix. 2021. URL: `https://www.statlect.com/fundamentals-of-probability/expected-value-and-Lebesgue-integral`.

[7] Alexander Drewitz. *Introduction to Probability and Statistics*. Preliminary version, February 1. University of Cologne, 2019.

[8] Peter Orbanz. *Functional Conjugacy in Parametric Bayesian Models*. Technical Report. University of Cambridge, 2009.

[9] Daniel V. Tausk. *A Basic Introduction to Probability and Statistics for Mathematicians*. Date: January 24th, 2023. 2023.

[10] Edward E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, 1978, p. 25.

[11] Glenn Shafer. "BELIEF FUNCTIONS AND POSSIBILITY MEASURES." English (US). In: *Anal of Fuzzy Inf.* CRC Press Inc, 1987, pp. 51–84. ISBN: 0849362962.

[12] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer, 2009. DOI: 10.1007/978-0-387-92407-6.

[13] E. T. Jaynes. "Probability Theory - The Logic of Science."

[14] E. T. Jaynes. "Prior Probabilities." In: *IEEE Transactions on Systems Science and Cybernetics* SSC-4 (1968), pp. 227–241.

[15] E. T. Jaynes. "Marginalization and Prior Probabilities." In: *Bayesian Analysis in Econometrics and Statistics*. Ed. by A. Zellner. Reprinted in [**jaynes_maximum_entropy_formalism**]. Amsterdam: North-Holland Publishing Company, 1980.

[16] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons, 1971.

[17] E. T. Jaynes. "Where Do We Stand On Maximum Entropy?" In: *The Maximum Entropy Formalism*. Ed. by R. D. Levine and M. Tribus. MIT Press, 1978, pp. 15–118.

[18] J. E. Shore and R. W. Johnson. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." In: *IEEE Transactions on Information Theory* IT-26.1 (1980), pp. 26–37.

[19] J. E. Shore and R. W. Johnson. "Properties of Cross-Entropy Minimization." In: *IEEE Transactions on Information Theory* IT-27.4 (1981), pp. 472–482.

[20] E. T. Jaynes. "Information Theory and Statistical Mechanics." In: *Phys. Rev.* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: http://prola.aps.org/abstract/PR/v106/i4/p620_1.

[21] J. NEYMAN and E. S. PEARSON. "ON THE USE AND INTERPRETATION OF CERTAIN TEST CRITERIA FOR PURPOSES OF STATISTICAL INFERENCE." In: *Biometrika* 20A.3-4 (Dec. 1928), pp. 263–294. ISSN: 0006-3444. DOI: 10.1093/biomet/20A.3-4.263. eprint: https://academic.oup.com/biomet/article-pdf/20A/3-4/263/1037410/20A-3-4-263.pdf. URL: https://doi.org/10.1093/biomet/20A.3-4.263.

[22] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[23] A. Wald. "Sequential Tests of Statistical Hypotheses." In: *The Annals of Mathematical Statistics* 16.2 (1945), pp. 117 –186. DOI: 10.1214/aoms/1177731118. URL: https://doi.org/10.1214/aoms/1177731118.

[24] Jerzy Neyman and Elizabeth Letitia Scott. "Consistent Estimates Based on Partially Consistent Observations." In: *Econometrica* 16 (1948), p. 1. URL: https://api.semanticscholar.org/CorpusID:155631889.

[25] E.L. Lehmann. *Testing Statistical Hypotheses*. Probability and Statistics Series. Wiley, 1986. ISBN: 9780471840831. URL: https://books.google.dk/books?id=jexQAAAAMAAJ.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[27] C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. 2nd. See Chapter 3 for the CramÃl'r-Rao inequality and its applications. New York: John Wiley & Sons, 1973. ISBN: 978-0-471-34969-5.

[28] T. Bayes. "An essay towards solving a problem in the doctrine of chances." In: *Phil. Trans. of the Royal Soc. of London* 53 (1763), pp. 370–418.

[29] Pierre-Simon Laplace. *ThÃl'orie analytique des probabilitÃl's*. Paris: Courcier, 1812. URL: `http://gallica.bnf.fr/ark:/12148/bpt6k88764q`.

[30] Bruno de Finetti. "La prévision : ses lois logiques, ses sources subjectives." fr. In: *Annales de l'institut Henri Poincaré* 7.1 (1937), pp. 1–68. URL: `http://www.numdam.org/item/AIHP_1937__7_1_1_0`.

[31] Harold Jeffreys. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. 1939. ISBN: 978-0-19-850368-2, 978-0-19-853193-7.

[32] L. Savage. *The Foundations of Statistics*. New York: Wiley, 1954.

[33] D. C. Plaut, S. J. Nowlan, and G. E. Hinton. *Experiments on learning back propagation*. Tech. rep. CMU–CS–86–126. Pittsburgh, PA: Carnegie–Mellon University, 1986.

[34] J. M Hammersleyand and D. C. Handscomb. *Monte Carlo Methods*. London, Methuen., 1964.

[35] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. "Hybrid Monte Carlo." In: *Phys. Lett. B* 195 (1987), pp. 216–222. DOI: `10.1016/0370-2693(87)91197-X`.

[36] Radford M. Neal. Berlin, Heidelberg: Springer-Verlag, 1996. ISBN: 0387947248.

[37] Radford M. Neal. "MCMC using Hamiltonian dynamics." In: (2012). cite arxiv:1206.1901. URL: `http://arxiv.org/abs/1206.1901`.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

[39] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262018029 0262018020. URL: https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618.

[41] Manfred Fischer and Petra Staufer-Steinnocher. "Optimization in an Error Backpropagation Neural Network Environment with a Performance Test on a Spectral Pattern Classification Problem." In: *Geographical Analysis* 31 (Jan. 1999), pp. 89–108. DOI: 10.1111/gean.1999.31.1.89.

[42] Renzo Sancisi. "The visible matter – dark matter coupling." In: *Symposium - International Astronomical Union* 220 (June 2004), p. 233. DOI: 10.1017/S0074180900183299.

[43] R.A. Swaters, R.H. Sanders, and S.S. McGaugh. "Testing Modified Newtonian Dynamics with Rotation Curves of Dwarf and Low Surface Brightness Galaxies." In: *Astrophys. J.* 718 (2010), pp. 380–391. DOI: 10.1088/0004-637X/718/1/380. arXiv: 1005.5456 [astro-ph.CO].

[44] R.A. Swaters, R. Sancisi, J.M. van der Hulst, and T.S. van Albada. "The Link between the Baryonic Mass Distribution and the Rotation Curve Shape." In: *Mon. Not. Roy. Astron. Soc.* 425 (2012), p. 2299. DOI: 10.1111/j.1365-2966.2012.21599.x. arXiv: 1207.2729 [astro-ph.CO].

[45] Benoît Famaey and Stacy S. McGaugh. "Modified Newtonian Dynamics (MOND): Observational Phenomenology and Relativistic Extensions." In: *Living Reviews in Relativity* 15.1, 10 (Sept. 2012), p. 10. DOI: 10.12942/lrr-2012-10. arXiv: 1112.3960 [astro-ph.CO].

[46] Jonas Petersen and Mads T Frandsen. “A method for discriminating between dark matter models and MOND modified inertia via galactic rotation curves.” In: *Monthly Notices of the Royal Astronomical Society* 496.2 (June 2020), pp. 1077–1091. ISSN: 0035-8711. DOI: `10.1093/mnras/staa1541`. eprint: `https://academic.oup.com/mnras/article-pdf/496/2/1077/33419277/staa1541.pdf`. URL: `https://doi.org/10.1093/mnras/staa1541`.

[47] James Binney and Scott Tremaine. *Galactic Dynamics: Second Edition*. 2. 2008.

[48] Federico Lelli, Stacy S. McGaugh, James M. Schombert, and Marcel S. Pawlowski. In: *Astrophys. J.* 836.2 (2017), p. 152. DOI: `10.3847/1538-4357/836/2/152`. arXiv: `1610.08981 [astro-ph.GA]`.

[49] Stacy McGaugh, Federico Lelli, and Jim Schombert. In: *Phys. Rev. Lett.* 117.20 (2016), p. 201101. DOI: `10.1103/PhysRevLett.117.201101`. arXiv: `1609.05917 [astro-ph.GA]`.

[50] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult.” In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166.

[51] Y. Bengio and P. Frasconi. “Diffusion of Context and Credit Information in Markovian Models.” In: *arXiv e-prints*, cs/9510101 (Sept. 1995), cs/9510101. arXiv: `cs/9510101 [cs.AI]`.

[52] Oriol Abril-Pla et al. “PyMC: a modern, and comprehensive probabilistic programming framework in Python.” In: *PeerJ Computer Science* 9 (Sept. 2023), e1516. ISSN: 2376-5992. DOI: `10.7717/peerj-cs.1516`. URL: `https://doi.org/10.7717/peerj-cs.1516`.

[53] J. Petersen. “The Missing MAss Problem on Galactic Scales.” PhD thesis.

[54] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. "Equation of State Calculations by Fast Computing Machines." In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: `10.1063/1.1699114`. URL: `http://link.aip.org/link/?JCP/21/1087/1`.

[55] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications." In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: `10.1093/biomet/57.1.97`. eprint: `http://biomet.oxfordjournals.org/cgi/reprint/57/1/97.pdf`.

[56] M. Betancourt and Mark Girolami. "Hamiltonian Monte Carlo for Hierarchical Models." In: (Dec. 2013). DOI: `10.1201/b18502-5`.

[57] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002.

[58] John Skilling. "Nested Sampling." In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. Ed. by Rainer Fischer, Roland Preuss, and Udo Von Toussaint. Vol. 735. American Institute of Physics Conference Series. Nov. 2004, pp. 395–405. DOI: `10.1063/1.1835238`.

[59] John Skilling. "Skilling, J.: Nested sampling for general Bayesian computation. Bayesian Anal. 1(4), 833-860." In: *Bayesian Analysis* 1 (Dec. 2006), pp. 833–860. DOI: `10.1214/06-BA127`.

# Index