INTRODUCTION TO STATISTICS

THEORY, METHODS, AND APPLICATIONS

JONAS PETERSEN

This page is intentionally left blank

Contents

CHAPTER 1

Preface

Statistics is a mathematical discipline that uses probability theory (which, in turn, requires 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 into probability questions through 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, which is 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 applied aspects of probability theory. Although probabilities are well defined (see Chapter 3), their interpretation is not specified beyond their mathematical definition. This ambiguity has given rise to two competing interpretations of probability, leading to two major branches of statistics: Frequentist and Bayesian statistics. This book aims to explain how these competing branches of statistics fit together, as well as to provide a non-exhaustive presentation of some of the methods within both branches.

1.1 ACKNOWLEDGEMENTS

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

CHAPTER 2

Introduction to Set Theory

Set theory is a fundamental branch of mathematical logic that underpins much of mathematics, including probability theory. At its core, set theory concerns the concept of a set – a collection of distinct objects or elements. This introduction explores the essential properties and operations of sets to lay the groundwork for the axiomatic development 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.

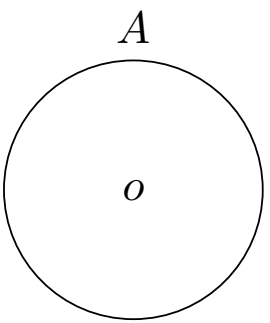

Figure 1: The graphical representation of a generic set A with generic elements o.

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.*

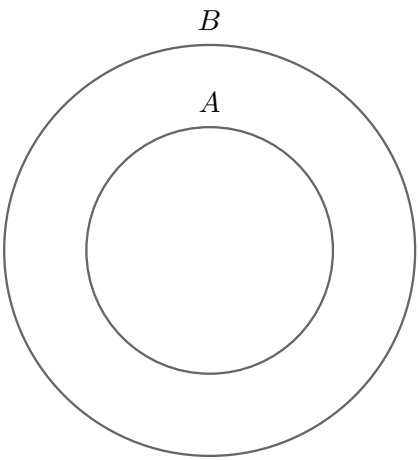

Figure 2: The graphical representation of $A \subseteq B$.

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 Ω, 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 $A \cup B$.*

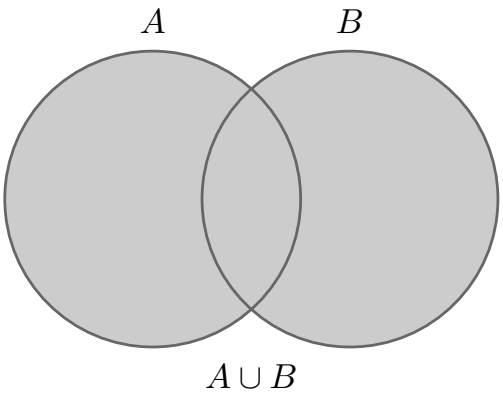

Figure 3: The figure show the union of sets A and B. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

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 $A \cap B$.*

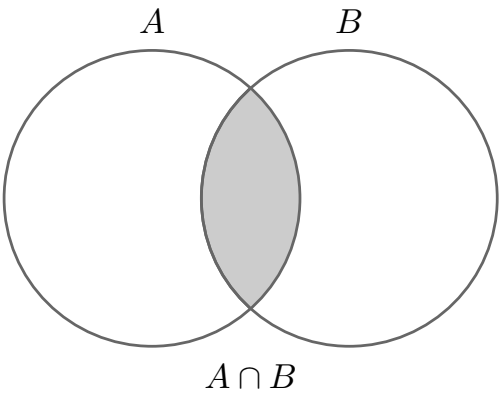

Figure 4: The figure show the intersection of sets A and B. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

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 $A \cap B = \emptyset$.*

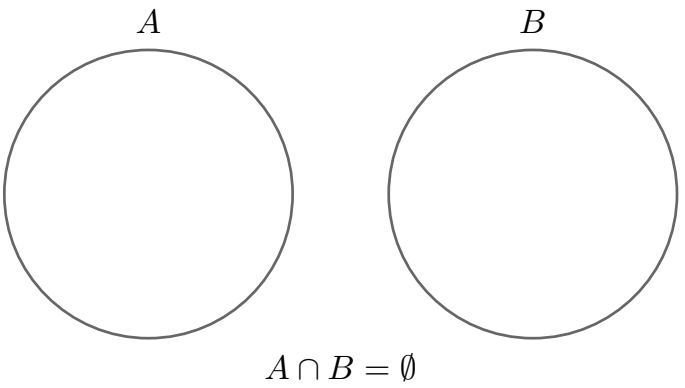

Figure 5: The figure show the case where the intersection of sets A and B is the empty set. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

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

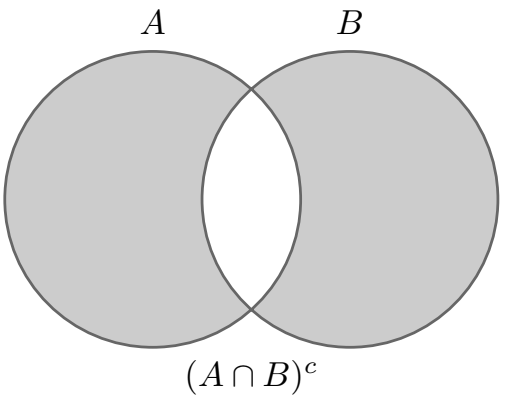

Figure 6: The figure show the complementary of the intersection of sets A and B. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

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 $B \setminus A$.*

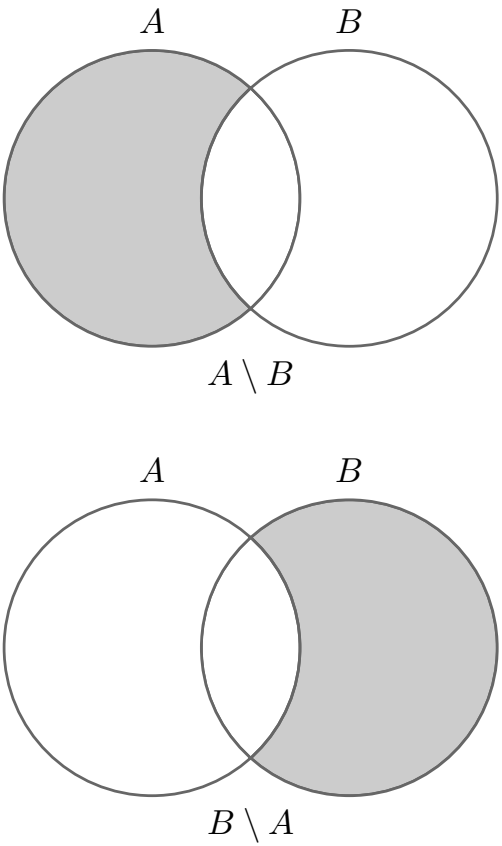

Figure 7: (top) show A minus B and (bottom) show B minus A. Each circle represent the sets and the colored region represent the result of the result of the binary operation.

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

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

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., $\bigcup_{i \in I} A_i = A$.*

A graphical representation of the set $A = \{A_1, A_2, A_3\}$, where A_j are partitions, is shown in Figure 8.

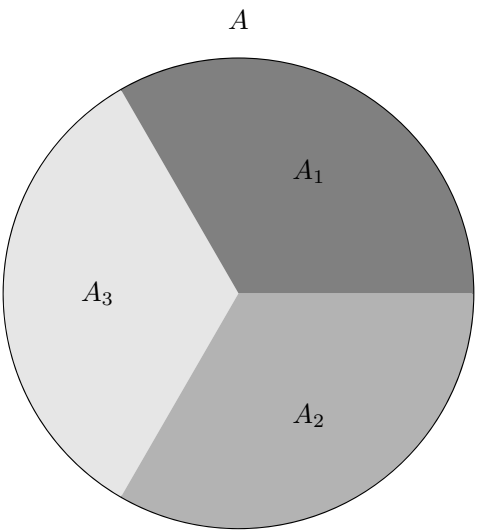

Figure 8: The figure show $A = \{A_1, A_2, A_3\}$ where A_j are partitions.

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.*

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 provides a mathematical framework for analyzing random experiments, where outcomes cannot be predicted with certainty in advance. Its objective is to systematically study and understand the possible outcomes of such experiments.

Definition 19 (Sample Space)**.** *The sample space, denoted by Ω, 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 $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. $E = \{\boxdot, \boxdot, \boxdot\}$, 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 σ-algebra satisfying the following properties:*

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

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

3. *Closure under countable union and intersection: If the events $E_1, E_2, \ldots$ are in the σ-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 σ-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 σ-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 $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ *and the event space (the set of all possible events) is given by*

$$\begin{aligned}\mathcal{F} &= \{\emptyset, \{\boxdot\}, \{\boxdot, \boxdot\}, \{\boxdot\}, \{\boxdot, \boxdot, \boxdot, \boxdot\}, \{\boxdot\}, \dots\} \\ &= 2^{\Omega}.\end{aligned} \tag{3}$$

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] Axiom 1-Axiom 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 Ω 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*

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

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 non-trivial 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 independent if*

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

Definition 26 (Conditional Independence). *Events E_1 and E_2 are conditionally independent given E_3 if*

$$\mathbb{P}(E_1, E_2|E_3) = \mathbb{P}(E_1|E_3)\mathbb{P}(E_2|E_3). \tag{10}$$

From Axiom 1-Axiom 3, Definition 8 and Definition 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*

$$\mathbb{P}(E_1, \ldots E_n) = \mathbb{P}(E_1) \prod_{j=2}^{n} \mathbb{P}(E_j|E_1, \ldots E_{j-1}). \tag{11}$$

Proof. From the definition of conditional probability in Equation 8

$$\mathbb{P}(E_1, E_2, \ldots, E_n) = \mathbb{P}(E_1|E_2, \ldots, E_n)\mathbb{P}(E_2, \ldots, E_n). \tag{12}$$

Using the definition of conditional probability again

$$\mathbb{P}(E_2, \ldots, E_n) = \mathbb{P}(E_2|\ldots, E_n)\mathbb{P}(\ldots, E_n). \tag{13}$$

Continuing in this way, Equation 11 follows. □

Equation 11 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*

$$\mathbb{P}(E_1|E_2, E_3) = \frac{\mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2, E_3)}. \tag{14}$$

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

$$\begin{aligned} \mathbb{P}(E_1, E_2, E_3) &= \mathbb{P}(E_1|E_2, E_3)\mathbb{P}(E_2, E_3) \\ &= \mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_3) \end{aligned} \tag{15}$$

from which

$$\mathbb{P}(E_1|E_2, E_3) = \frac{\mathbb{P}(E_2|E_1, E_3)\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2, E_3)} \tag{16}$$

which is Bayes theorem. □

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

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

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 Ω such that $\cup_i E_i = \Omega$. For mutually exclusive events (which a partition by definition is), finite additivity can be used such that

$$\sum_i \mathbb{P}(A, E_i) = \mathbb{P}\Big(\bigcup_i (A, E_i)\Big). \tag{18}$$

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

$$\begin{aligned} \bigcup_i (A, E_i) &= (A, \bigcup_i E_i) \\ &= (A, \Omega) \\ &= A. \end{aligned} \tag{19}$$

Combining Equation 18-Equation 19 then yields

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, E_i). \tag{20}$$

□

Example 3.3.

Consider the roll of a fair six-sided die. The sample space for this experiment is given by $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. *Let* $E_1 = \{\boxdot, \boxdot, \boxdot\}$ *and* $E_2 = \{\boxdot\}$ *be two events, then from Equation 8*

$$\begin{aligned} \mathbb{P}(E_1|E_2) &= \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_2)} \\ &= 1 \end{aligned} \tag{21}$$

where $\mathbb{P}(E_1, E_2) = \frac{1}{6}$ *since* $E_1, E_2 = E_1 \cap E_2 = E_2 = \{\boxdot\}$ *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 27 (Random Variable)**.** *A random variable* X *is a function*

$$X : \Omega \mapsto \Omega_X \tag{22}$$

that maps outcomes from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$, *where* Ω_X *is the codomain of* X *and* $\mathcal{F}_X$ *is a* σ*-algebra on* Ω_X. *The*

σ-algebra $\mathcal{F}_X$ ensures that X is measurable, meaning that for any set $B \in \mathcal{F}_X$, the preimage $X^{-1}(B)$ must belong to $\mathcal{F}$. Formally, this can be written as

$$X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{F}_X. \tag{23}$$

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 28 (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]*

$$\mathbb{P} \circ X^{-1} : \mathcal{F}_X \mapsto [0, 1] \tag{24}$$

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

Definition 29 (Expected value via image measure)**.** *Let X be a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let*

$$\mathbb{P}_X = \mathbb{P} \circ X^{-1} \tag{25}$$

be the image (pushforward) measure (Definition 28) of X on $(\Omega_X, \mathcal{F}_X)$. The expected value of X, denoted by $\mathbb{E}[X]$, can be defined viz

$$\mathbb{E}[X] \equiv \int_{\Omega_X} x \mathbb{P}_X(dx). \tag{26}$$

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

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

Theorem 6 (Law of the Unconscious Statistician)**.** *Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $g : \Omega_X \to \mathbb{R}$ be any measurable function. Denote the image (pushforward) measure (Definition 28) of X by $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. Then*

$$\mathbb{E}[g(X)] \equiv \int_{\Omega_X} g(x) \mathbb{P}_X(dx). \tag{27}$$

Definition 30 (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 $\mathrm{Var}[X]$, is defined viz*

$$\begin{aligned}\mathrm{Var}[X] &\equiv \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned} \tag{28}$$

Theorem 7 (Markov's Inequality)**.** *Let X be a non-negative random variable and $a > 0$. Then*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{29}$$

Proof. Let $1_{\{X \geq a\}}$ denote the indicator of the event $\{X \geq a\}$. Since X is non-negative and $a > 0$

$$a 1_{\{X \geq a\}} \leq X. \tag{30}$$

Taking expectations and using Theorem 5

$$a\,\mathbb{E}\left[1_{\{X \geq a\}}\right] \leq \mathbb{E}[X]. \tag{31}$$

But $\mathbb{E}[1_{\{X \geq a\}}] = \mathbb{P}(X \geq a)$, so

$$a\,\mathbb{P}(X \geq a) \leq \mathbb{E}[X], \tag{32}$$

and rearranging yields Markov's inequality

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{33}$$

□

Definition 31 (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 (Definition 28) is defined as the probability mass function*

$$\begin{aligned}p(X = x) &\equiv \mathbb{P}_X(\{x\}) \\ &= \mathbb{P}(X^{-1}(\{x\})).\end{aligned} \tag{34}$$

According to Axiom 1-Axiom 3

$$\sum_{x \in \Omega_X} p(X = x) = 1 \tag{35}$$

and

$$p(X = x) \geq 0, \quad \forall x \in \Omega_X. \tag{36}$$

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*

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} xp(X = x). \tag{37}$$

Definition 32 (Probability Density Function)**.** *Let $X : \Omega \to \Omega_X$ be a continuous random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $(\Omega_X, \mathcal{F}_X)$ be a measurable space. Denote the image (pushforward) measure (Definition 28) of X by*

$$\mathbb{P}_X(B) \equiv \mathbb{P}(X^{-1}(B)), \quad \forall B \in \mathcal{F}_X. \tag{38}$$

If there exists a non-negative measurable function

$$f : \Omega_X \to \mathbb{R}_{\geq 0} \tag{39}$$

such that

$$\mathbb{P}_X(B) = \int_B d\lambda(x) f(x), \quad \forall B \in \mathcal{B}(\Omega_X), \tag{40}$$

where λ denotes the Lebesgue measure and $\mathcal{B}(\Omega_X)$ is the Borel σ-algebra on Ω_X, then f is called the probability density function (PDF) of X. The PDF satisfies

$$f(x) \geq 0 \quad \forall x \in \Omega_X, \qquad \int_{\Omega_X} d\lambda(x) f(x) = 1. \tag{41}$$

Example 3.4.

Let X be a continuous random variable with PDF f. For an interval $[a, b] \subseteq \Omega_X$,

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}_X([a, b]) \\ &= \int_{[a,b]} d\lambda(x) f(x) \\ &= \int_a^b dx f(x). \end{aligned} \tag{42}$$

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*

$$\mathbb{E}[X] = \int_{\Omega_X} dx x f(x). \tag{43}$$

Theorem 10 (Total expectation)**.** *Let $X : \Omega \mapsto \Omega_X$ and $Y : \Omega \mapsto \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]], \tag{44}$$

where the subscript indicates the probability distribution the expectation is taken with respect to.

Proof.

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} dx x f(X = x) \\
&= \int_{\Omega_Y} dy \int_{\Omega_X} dx x f(X = x, Y = y) \\
&= \int_{\Omega_Y} dy f(Y = y) \int_{\Omega_X} dx x f(X = x|Y = y) \\
&= \int_{\Omega_Y} dy f(Y = y) \underbrace{\int_{\Omega_X} dx x f(X = x|Y = y)}_{=\mathbb{E}_{X|Y}[X|Y]} \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]].
\end{aligned} \tag{45}$$

□

Theorem 11 (Expectation of product of independent random variables)**.** *Let $X : \Omega \mapsto \Omega_X$ and $Y : \Omega \mapsto \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that*

$$f(X = x, Y = y) = f(X = x) f(Y = y), \tag{46}$$

then $\mathbb{E}_{XY}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]$.

Proof.

$$\begin{aligned}
\mathbb{E}_{XY}[XY] &= \int_{\Omega_X} dx \int_{\Omega_Y} dy x y f(X = x, Y = y) \\
&= \int_{\Omega_X} dx x f(X = x) \int_{\Omega_Y} dy y f(Y = y) \\
&= \mathbb{E}_X[X]\mathbb{E}_Y[Y]
\end{aligned} \tag{47}$$

□

Definition 33 (Covariance). *Let $X : \Omega \mapsto \Omega_X$ and $Y : \Omega \mapsto \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the covariance of X and Y, denoted by $\mathrm{Cov}[X, Y]$, is defined viz*

$$\begin{aligned} \mathrm{Cov}[X, Y] &= \mathbb{E}_{XY}[(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])] \\ &= \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y], \end{aligned} \tag{48}$$

Theorem 12 (Covariance of independent random variables). *Let $X : \Omega \to \Omega_X$ and $Y : \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If X and Y are independent, then their covariance is*

$$\mathrm{Cov}[X, Y] = 0. \tag{49}$$

Proof. Using $\mathbb{E}_{XY}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]$ (Theorem 11) in Definition 33 yield $\mathrm{Cov}[X, Y] = 0$. □

Definition 34 (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 $\mathrm{Corr}[X, Y]$, is defined as*

$$\begin{aligned} \mathrm{Corr}[X, Y] &= \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} \\ &= \frac{\mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2)\,(\mathbb{E}_Y[Y^2] - \mathbb{E}_Y[Y]^2)}}. \end{aligned} \tag{50}$$

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 35 (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]*

$$f(Y = y) = f(X = g^{-1}(y)) \left| \frac{d}{dY} \left(g^{-1}(Y)\right) \right|_{Y=y}. \tag{51}$$

Example 3.5.

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

$$g^{-1}(y) = \frac{y - b}{a} \tag{52}$$

Using Definition 35

$$\begin{aligned} f(Y = y) &= f\left(X = g^{-1}(y)\right) \left| \frac{d}{dY} \left(g^{-1}(Y)\right) \right|_{Y=y} \\ &= f\left(X = \frac{y-b}{a}\right) \left| \frac{d}{dY} \left(\frac{Y-b}{a}\right) \right|_{Y=y} \\ &= f\left(X = \frac{y-b}{a}\right) \left| \frac{1}{a} \right|. \end{aligned} \tag{53}$$

Thus, the PDF of Y is

$$f(Y = y) = \frac{1}{|a|} f\left(X = \frac{y-b}{a}\right). \tag{54}$$

Example 3.6.

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

$$g^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{55}$$

Using Definition 35

$$\begin{aligned} f(Y=y) &= f\left(X=g^{-1}(y)\right)\left|\frac{d}{dY}\left(g^{-1}(Y)\right)\right|_{Y=y} \\ &= \text{const}\cdot\frac{1-Y}{Y}\left(\frac{1}{1-Y}+\frac{Y}{(1-Y)^2}\right)\Bigg|_{Y=y} \\ &= \text{const}\cdot Y^{-1}(1-Y)^{-1}|_{Y=y} \\ &= \text{Beta}(Y=y|a=0,b=0). \end{aligned} \tag{56}$$

Definition 36 (Error Propagation)**.** *Let $X_1,\dots,X_n$ be continuous random variables defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, and let $g:\mathbb{R}^n\to\mathbb{R}$ be a differentiable function of these variables. The variance of g, which quantifies the uncertainty in g due to the uncertainties in the random variables $X_1,\dots,X_n$, can be written*

$$\begin{aligned} \text{Var}[g] &= \mathbb{E}\left[(g-\mathbb{E}[g])^2\right] \\ &= \sum_{i=1}^{n}\left.\frac{\partial g}{\partial X_i}\right|^2_{X=\mathbb{E}[X]}\text{Var}[X_i] + \sum_{i\neq j}\left.\frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}\right|_{X=\mathbb{E}[X]}\text{Cov}[X_i,X_j] \\ &\quad + \mathcal{O}(\|X-\mathbb{E}[X]\|^3). \end{aligned} \tag{57}$$

Proof. With the notation $X=(X_1,\dots,X_n)$ and $\mathbb{E}[X]=(\mathbb{E}[X_1],\dots,\mathbb{E}[X_n])$ $g(X)$ can be written as a Taylor expansion around $\mathbb{E}[X]$ viz

$$g(X) = g(\mathbb{E}[X]) + \sum_{i=1}^{n}\left.\frac{\partial g}{\partial X_i}\right|_{X=\mathbb{E}[X]}(X_i-\mathbb{E}[X_i]) + \mathcal{O}(\|X-\mathbb{E}[X]\|^2). \tag{58}$$

Consequently

$$\begin{aligned} \mathbb{E}[g(X)] &= g(\mathbb{E}[X]) + \sum_{i=1}^{n}\left.\frac{\partial g}{\partial X_i}\right|_{X=\mathbb{E}[X]}\underbrace{\mathbb{E}[X_i-\mathbb{E}[X_i]]}_{\to 0} + \mathcal{O}(\|X-\mathbb{E}[X]\|^2) \\ &= g(\mathbb{E}[X]) + \mathcal{O}(\|X-\mathbb{E}[X]\|^2) \end{aligned} \tag{59}$$

meaning the variance of g can be approximated viz

$$\begin{aligned}\operatorname{Var}[g] &= \mathbb{E}\big[(g-\mathbb{E}[g])^2\big] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])\frac{\partial g}{\partial X_i}\bigg|_{X=\mathbb{E}[X]}+\mathcal{O}(\|X-\mathbb{E}[X]\|^2)\right)^2\right] \\ &= \sum_{i=1}^{n}\frac{\partial g}{\partial X_i}\bigg|_{X=\mathbb{E}[X]}^2\operatorname{Var}[X_i]+\sum_{i\neq j}\frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}\bigg|_{X=\mathbb{E}[X]}\operatorname{Cov}[X_i,X_j] \\ &\quad+\mathcal{O}(\|X-\mathbb{E}[X]\|^3).\end{aligned} \tag{60}$$

□

Remark 1 (Independent Random Variables)**.** *If the random variables $X_1,\dots,X_n$ are independent, then $\operatorname{Cov}[X_i,X_j]=0$ for all $i\neq j$ (see Theorem 12). In this case, Definition 36 simplifies to*

$$\operatorname{Var}[g(X_1,\dots,X_n)]\approx\sum_{i=1}^{n}\left(\frac{\partial g}{\partial X_i}\bigg|_{X=\mathbb{E}[X]}\right)^2\operatorname{Var}[X_i]. \tag{61}$$

This is the commonly used form of the linear error-propagation formula for independent variables.

Example 3.7.

A company produce square plates. Let the plate dimensions be characterized by two independent random variables $X\sim\operatorname{Norm}(2m,(0.01m)^2)$ and $Y\sim\operatorname{Norm}(3m,(0.02m)^2)$ and the area given by XY. Determine the variance of XY. From Definition 36, the exact variance is

$$\begin{aligned}\operatorname{Var}[XY] &= \mathbb{E}[(XY)^2]-(\mathbb{E}[XY])^2 \\ &= \Big(\operatorname{Var}[X]+\mathbb{E}[X]\Big)\Big(\operatorname{Var}[Y]+\mathbb{E}[Y]\Big)-\mathbb{E}[X]^2\mathbb{E}[Y]^2 \\ &= \mathbb{E}[Y]^2\operatorname{Var}[X]+\mathbb{E}[X]^2\operatorname{Var}[Y]+\operatorname{Var}[X]\operatorname{Var}[Y]\end{aligned} \tag{62}$$

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 from Remark 1

$$\begin{aligned}\operatorname{Var}[XY] &\approx \sum_{i=X,Y}\left(\frac{\partial(XY)}{\partial i}\bigg|_{X=\mathbb{E}[X],Y=\mathbb{E}[Y]}\right)^2\operatorname{Var}[i] \\ &= \mathbb{E}[Y]^2\operatorname{Var}[X]+\mathbb{E}[X]^2\operatorname{Var}[Y]\end{aligned} \tag{63}$$

Comparing Equation 62 and Equation 63 the relative difference can be written

$$\frac{\mathrm{Var}[XY] - \mathrm{Var}[XY]|_{linear\ approximation}}{\mathrm{Var}[XY]} = \frac{\mathrm{Var}[X]\,\mathrm{Var}[Y]}{\mathrm{Var}[XY]} \simeq 1.6 \cdot 10^{-5}. \tag{64}$$

Example 3.8.

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

$$\begin{aligned} \Omega_{child\ 1} &= \{\text{👦},\text{👧}\}, \\ \Omega_{child2} &= \{\text{👦},\text{👧}\} \end{aligned} \tag{65}$$

such that

$$\begin{aligned} \Omega &= \Omega_{child\ 1} \times \Omega_{child\ 2} \\ &= \{(\text{👦},\text{👦}), (\text{👦},\text{👧}), (\text{👧},\text{👦}), (\text{👧},\text{👧})\}. \end{aligned} \tag{66}$$

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 Definition 31, such that e.g.

$$p(B = 1, G = 1) = \mathbb{P}(\{(\text{👦},\text{👧})\}). \tag{67}$$

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"

   $$p(G = 1, B \geq 1) = \frac{p(B \geq 1|G = 1)p(G = 1)}{p(B \geq 1)}. \tag{68}$$

   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.

   $$p(B \geq 1) = 1 - p(G = 2, B = 0) = \frac{3}{4}, \tag{69}$$

so

$$p(G = 1|B \geq 1) = \frac{2}{3}. \tag{70}$$

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

$$p(G = 1|B = 1) = \frac{1}{2}. \tag{71}$$

Example 3.9.

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 that is present in 1% *of the population. Let* E *denote the event that a person has the blood type found at the crime scene. Then*

$$\mathbb{P}(E) = 0.01. \tag{72}$$

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 prosecutor's fallacy. What is wrong with this argument?*

The prosecutor's claim can be written as

$$\mathbb{P}(E \mid \text{innocent}) = 0.01 \Rightarrow \mathbb{P}(\text{guilty} \mid E) = 0.99. \tag{73}$$

To investigate this claim, note that from Definition 24,

$$\begin{aligned} \mathbb{P}(E \mid \text{innocent}) &= \frac{\mathbb{P}(E, \text{innocent})}{\mathbb{P}(\text{innocent})} \\ &= \frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})}\mathbb{P}(E). \end{aligned} \tag{74}$$

Hence, in general, $\mathbb{P}(E \mid \text{innocent}) \neq \mathbb{P}(E)$. *Suppose there are* N *people in the world, and* $M \leq N$ *of these have the blood type found at the crime scene. In that case,*

$$\frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})} = \frac{\frac{M-1}{M}}{\frac{N-1}{N}}, \tag{75}$$

which approaches 1 *in the limit* $N \to \infty$. *Hence,* $\mathbb{P}(E \mid \text{innocent}) \simeq \mathbb{P}(E)$ *can be a good approximation, but it is not an exact relation.*
Assuming $\mathbb{P}(E \mid \text{innocent}) = 0.01$, *the prosecutor's claim can be further analyzed using Definition 24, viz.*

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(\text{guilty}, E) + \mathbb{P}(\text{innocent}, E)}{\mathbb{P}(E)}. \quad (76)$$

Innocent and guilty are complementary events that form a partition of the sample space, meaning (Theorem 3)

$$\mathbb{P}(\text{guilty}, E) + \mathbb{P}(\text{innocent}, E) = \mathbb{P}(E), \quad (77)$$

and thereby

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = 1. \quad (78)$$

This means that if $\mathbb{P}(\text{guilty} \mid E) = 0.99$, *then* $\mathbb{P}(\text{innocent} \mid E) = 0.01$, *and from Theorem 2,*

$$\mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(E \mid \text{innocent})\,\mathbb{P}(\text{innocent})}{\mathbb{P}(E)} \quad (79)$$

From equation (79), *it is clear that in general*

$$\mathbb{P}(E \mid \text{innocent}) \neq \mathbb{P}(\text{innocent} \mid E), \quad (80)$$

and so even if $\mathbb{P}(E \mid \text{innocent}) = 0.01$, *the prosecutor's claim (Equation 73) is not true.*

Example 3.10.

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

$$\begin{aligned}\text{Var}[X+Y] &= \mathbb{E}_{XY}[(X+Y-\mathbb{E}_{XY}[X+Y])^2] \\ &= \mathbb{E}_X[(X-\mathbb{E}_X[X])^2] + \mathbb{E}_Y[(Y-\mathbb{E}_Y[Y])^2] \\ &\quad + 2\mathbb{E}_{XY}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}[X,Y].\end{aligned} \quad (81)$$

Example 3.11.

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

$$\begin{aligned}\mathbb{P}(s|p) &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p)} \\ &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p|s)\mathbb{P}(s)+\mathbb{P}(p|h)\mathbb{P}(h)}\end{aligned} \tag{82}$$

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*

$$\mathbb{P}(s|p) \simeq 0.0098. \tag{83}$$

Example 3.12.

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*

$$z^* = \underset{z}{\operatorname{argmax}}(\mathbb{P}(z|o_3, c_1)). \tag{84}$$

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

$$\mathbb{P}(z_1|o_3, c_1) = \frac{\mathbb{P}(o_3|c_1, z_1)\mathbb{P}(c_1, z_1)}{\mathbb{P}(o_3, c_1)} \tag{85}$$

with

$$\begin{aligned}\mathbb{P}(o_3, c_1) &= \sum_i \mathbb{P}(o_3, c_1, z_i) \\ &= \mathbb{P}(o_3, c_1, z_1) + \mathbb{P}(o_3, c_1, z_2) + \mathbb{P}(o_3, c_1, z_3) \\ &= \mathbb{P}(o_3|c_1, z_1)\mathbb{P}(c_1, z_1) + \mathbb{P}(o_3|c_1, z_2)\mathbb{P}(c_1, z_2) \\ &\quad + \mathbb{P}(o_3|c_1, z_3)\mathbb{P}(c_1, z_3).\end{aligned} \tag{86}$$

$\mathbb{P}(o_3|c_1, z_3) = 0$ *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*

$$\begin{aligned}\mathbb{P}(z_1|o_3, c_1) &= \frac{\mathbb{P}(o_3|c_1, z_1)}{\sum_i \mathbb{P}(o_3|c_1, z_i)} \\ &= \frac{1}{3}.\end{aligned} \tag{87}$$

Similarly

$$\begin{aligned}\mathbb{P}(z_2|o_3, c_1) &= \frac{\mathbb{P}(o_3|c_1, z_2)}{\sum_i \mathbb{P}(o_3|c_1, z_i)} \\ &= \frac{2}{3}.\end{aligned} \tag{88}$$

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 3.13.

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

$$\begin{aligned} \text{Corr}[X, Y] &= \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\,\text{Var}[Y]}} \\ &= \frac{\mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{\text{Var}[X]\,\text{Var}[Y]}} \end{aligned} \tag{89}$$

In this case for the nominator

$$\begin{aligned} \text{Cov}[X, Y] &= \int dx x^3 p(x) - \int dx' x' p(x') \int dx'' x''^2 p(x'') \\ &= \frac{1}{b-a}\int_a^b x^3 dx - \frac{1}{(b-a)^2}\int_a^b dx' x' \int_a^b dx'' x''^2 \\ &= \frac{1}{12}(a-b)^2(a+b) \\ &= 0 \end{aligned} \tag{90}$$

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*

$$\begin{aligned} \text{Var}[X]\,\text{Var}[X^2] &= \big(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2\big)\big(\mathbb{E}_X[X^4] - \mathbb{E}_X[X^2]^2\big) \\ &= \frac{1}{540}(b-a)^4(4a^2 + 7ab + 4b^2) \\ &\neq 0. \end{aligned} \tag{91}$$

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

Example 3.14.

Let $X \sim \text{Norm}(\mu = 0, \sigma^2 = 1)$ *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* $Y \sim \text{Norm}(\mu = 0, \sigma^2 = 1)$*.*

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

$$\begin{aligned} \mathbb{E}_Y[Y] &= \mathbb{E}_{WX}[WX] \\ &= \mathbb{E}_W[W]\cancelto{0}{\mathbb{E}_X[X]} \\ &= 0. \end{aligned} \tag{92}$$

The variance

$$\begin{aligned}\mathrm{Var}[Y] &= \mathbb{E}_Y[Y^2] - \cancelto{0}{\mathbb{E}_Y[Y]^2} \\ &= \mathbb{E}_{WX}[W^2X^2] \\ &= \mathbb{E}_W[W^2]\mathbb{E}_X[X^2] \\ &= \mathbb{E}_W[W^2]\,\mathrm{Var}[X]\end{aligned} \tag{93}$$

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

$$\begin{aligned}\mathbb{E}_W[W^2] &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 p(W = w_i) \\ &= \frac{1}{2}[(-1)^2\frac{1}{2} + 1^2\frac{1}{2}] \\ &= 1\end{aligned} \tag{94}$$

so $\mathrm{Var}[Y] = 1$.

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

$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathrm{Cov}[X, WX] \\ &= \mathbb{E}_{WX}[WX^2] - \mathbb{E}_X[X]\mathbb{E}_{WX}[WX] \\ &= \mathbb{E}_W[W]\mathbb{E}_X[X^2] - \mathbb{E}_W[W]\mathbb{E}_X[X]^2 \\ &= \mathbb{E}_W[W]\,\mathrm{Var}[X] \\ &= 0\end{aligned} \tag{95}$$

where for the last equality it has been used that

$$\begin{aligned}\mathbb{E}_W[W] &= \frac{1}{n}\sum_{i=1}^{n} w_i p(W = w_i) \\ &= \frac{1}{2}[(-1)\frac{1}{2} + 1\frac{1}{2}] \\ &= 0\end{aligned} \tag{96}$$

Example 3.15.
Prove that $-1 \leq \mathrm{Corr}[X,Y] \leq 1$.

Since the variance is defined as positive definite

$$\begin{aligned} 0 &\leq \mathrm{Var}\left[\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right] \\ &= \frac{\mathrm{Var}[X]}{\sigma_X^2} + \frac{\mathrm{Var}[Y]}{\sigma_Y^2} \pm \frac{2}{\sigma_X \sigma_Y}\,\mathrm{Cov}[X,Y] \\ &= \frac{\mathrm{Var}[X]}{\sigma_X^2} + \frac{\mathrm{Var}[Y]}{\sigma_Y^2} \pm 2\,\mathrm{Corr}[X,Y] \\ &= 2 \pm 2\,\mathrm{Corr}[X,Y] \end{aligned} \tag{97}$$

where for the last equality it has been used that $\sigma_i^2 = \mathrm{Var}[i]$. *From equation* (97) *the result follows*

$$-1 \leq \mathrm{Corr}[X,Y] \leq 1. \tag{98}$$

Example 3.16.
Show that if $Y = aX + b$ *for some parameters* $a > 0$ *and* b, *then* $\mathrm{Corr}[X,Y] = 1$. *Similarly show that if* $a < 0$, *then* $\mathrm{Corr}[X,Y] = -1$.

$$\mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} \tag{99}$$

$$\begin{aligned} \mathrm{Cov}[X,Y] &= \mathbb{E}_{XY}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\ &= \mathbb{E}_X[X(aX+b)] - \mathbb{E}_X[X]\mathbb{E}_X[aX+b] \\ &= a\mathbb{E}_X[X^2] + b\mathbb{E}_X[X] - a\mathbb{E}_X[X]^2 - b\mathbb{E}_X[X] \\ &= a\,\mathrm{Var}[X] \end{aligned} \tag{100}$$

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{Var}[aX+b] \\ &= a^2\,\mathrm{Var}[X] + \cancelto{0}{\mathrm{Var}[b]} + \cancelto{0}{2\mathrm{Cov}[aX,b]} \\ &= a^2\,\mathrm{Var}[X] \end{aligned} \tag{101}$$

$$\begin{aligned} \mathrm{Corr}[X,Y] &= \frac{a\,\mathrm{Var}[X]}{\sqrt{a^2\,\mathrm{Var}[X]\,\mathrm{Var}[X]}} \\ &= \frac{a}{|a|} \end{aligned} \tag{102}$$

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

Example 3.17.

Let n *denote the total number of candidates, presented sequentially in uniformly random order. After observing the* k*-th candidate, an irrevocable decision must be made: either to accept or reject the candidate. The objective is to maximize the probability of selecting the candidate with the highest rank among all* n *candidates.*

Let $r \in \{0, 1, \ldots, n-1\}$ *denote the number of candidates to be automatically rejected. Define the strategy* σ_r *as follows: reject the first* r *candidates, then select the first subsequent candidate whose observed rank exceeds all ranks observed among the first* r *candidates. Let* $\mathbb{P}_n(r)$ *denote the probability that strategy* σ_r *selects the best candidate. For* $k \in \{r+1, \ldots, n\}$*, the probability that the* k*-th candidate is the best and is selected by* σ_r *is*

$$\mathbb{P}(\textit{best at position } k \textit{ and selected}) = \frac{1}{n} \cdot \frac{r}{k-1}. \tag{103}$$

Summation over all admissible positions yields

$$\begin{aligned} \mathbb{P}_n(r) &= \sum_{k=r+1}^{n} \frac{r}{n(k-1)} \\ &= \frac{r}{n} \sum_{k=r}^{n-1} \frac{1}{k}. \end{aligned} \tag{104}$$

For large n*, the sum may be approximated by an integral:*

$$\begin{aligned} \mathbb{P}_n(r) &\approx \frac{r}{n} \int_r^n \frac{dx}{x} \\ &= \frac{r}{n} \ln \frac{n}{r}. \end{aligned} \tag{105}$$

Setting $r = \alpha n$*,* $\alpha \in (0,1)$*, gives*

$$\mathbb{P}_n(\alpha n) \approx \alpha \ln \frac{1}{\alpha}. \tag{106}$$

The maximum occurs at

$$\begin{aligned} \frac{d}{d\alpha}\left(\alpha \ln \tfrac{1}{\alpha}\right) &= \ln \frac{1}{\alpha} - 1 \\ &= 0 \end{aligned} \tag{107}$$

meaning

$$\alpha = \frac{1}{e}. \tag{108}$$

The optimal stopping strategy consists of rejecting the first $\frac{n}{e}$ *candidates, then selecting the first candidate superior to all previously observed. The maximum probability of success converges to* $\mathbb{P}_n \to \frac{1}{e} \approx 0.368$ *as* $n \to \infty$*. This strategy extends naturally to situations where candidates arrive sequentially over time. In the continuous-time setting, the first* $1/e$ *fraction of the time interval is used purely for observation, and thereafter the first candidate exceeding all previous observations is selected.*

CHAPTER 4

Assigning Probability Functions

The axioms and definitions (Axiom 1-Axiom 3, Definition 8 and Definition 9) of probability theory 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 [8]. Logical analysis may be used to exploit the group invariances of a problem [9]. Logical analysis may be used to ensure consistency when uninteresting or nuisance parameter are marginalized from probability functions [10]. And last, logical analysis may be applied in the form of the principle of maximum entropy to yield probability functions [9, 11–14]. Of these techniques the principle of maximum entropy is probably the most powerful.

4.1 THE PRINCIPLE OF MAXIMUM ENTROPY

The principle of maximum entropy, first proposed by Jaynes [15], addresses the problem of assigning a probability distribution to a random variable in a way that is maximally noncommittal with respect to missing information. Let Z be a generic random variable describing an abstract experiment. Its probability distribution is denoted by $p(z|\lambda, I)$, parameterized by $\lambda = \{\lambda_0, \ldots, \lambda_n\}$. Here, I represents the background information related to the system (Definition 37).

Definition 37 (Background information)**.** *Background information, denoted by I, consists of all prior knowledge, assumptions, and constraints that are available before observing the outcome of a random experiment. This includes, but is not limited to:*

1. *Known properties of the system or phenomenon being modeled, such as symmetries, invariances, or physical laws.*
2. *Knowledge of which probability distributions or families of distributions are plausible for the random variables.*
3. *Preferences, biases, or prior beliefs regarding particular modeling methods, distributions, or parameter choices.*
4. *Any additional constraints, such as known moments, support, or relationships between variables.*

Background information formally determines the class of admissible probability distributions and methods considered suitable for representing uncertainty in the system.

The maximum entropy principle asserts that the probability distribution $p(z|\lambda, I)$ that best represents the current state of knowledge is the one that maximizes the constrained entropy [1]. Using the method of Lagrange multipliers, the constrained entropy is expressed via the Lagrangian

$$\mathcal{L} = \int dz F, \tag{109}$$

with

$$F = -p(z|\lambda, I) \ln \frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \sum_{j=1}^{n} \lambda_j C_j(z). \tag{110}$$

Here, $m(z)$, referred to as the reference measure or Lebesgue measure, ensures that the entropy

$$-\int dz p(z|\lambda, I) \ln \frac{p(z|\lambda, I)}{m(z)} \tag{111}$$

is invariant under changes of variables. The functions $C_j(z)$ represent additional constraints imposed by the background information I (Definition 37), beyond the normalization constraint. The Lagrangian is maximized by solving the Euler-Lagrange equation

$$\frac{\partial F}{\partial p(z|\lambda, I)} - \frac{d}{dz} \frac{\partial F}{\partial p'(z|\lambda, I)} = 0, \tag{112}$$

where $p'(z|\lambda, I) = \partial p(z|\lambda, I)/\partial z$. Since F does not depend on derivatives of p, the equation simplifies to

$$\frac{\partial F}{\partial p(z|\lambda, I)} = 0. \tag{113}$$

From this, the maximum entropy distribution is obtained as

$$\frac{\partial F}{\partial p(z|\lambda, I)} = -\ln \frac{p(z|\lambda, I)}{m(z)} - 1 - \sum_j \lambda_j C_j(z) = 0, \tag{114}$$

leading to

$$p(z|\lambda, I) = m(z)e^{-1-\sum_j \lambda_j C_j(z)} = \tilde{m}(z)e^{-\sum_j \lambda_j C_j(z)}, \tag{115}$$

where $\tilde{m}(z) \equiv m(z)e^{-1}$. Enforcing normalization $\int dz p(z|\lambda, I) = 1$ gives

$$p(z|\lambda, I) = \frac{\tilde{m}(z)e^{-\sum_j \lambda_j C_j(z)}}{\int dz' \tilde{m}(z')e^{-\sum_j \lambda_j C_j(z')}}. \tag{116}$$

The reference measure m is invariant under parameter transformations, and the Lagrange multipliers λ_j are determined by the additional constraints, e.g., on the mean or variance of Z.

Example 4.1.

Consider a random variable, Z, with unlimited support, $z \in [-\infty, \infty]$, assumed to be symmetric around a single peak defined by the mean μ, standard deviation σ. In this case $\lambda = \{\lambda_0, \lambda_1, \lambda_2\}$, where it will be shown that λ_1, λ_2 are related to μ, σ. In this case F can be written

$$\begin{aligned} F = &- p(z|\lambda, I) \ln \left(\frac{p(z|\lambda, I)}{m(z)} \right) - \lambda_0 p(z|\lambda, I) \\ &- \lambda_1 p(z|\lambda, I) z - \lambda_2 p(z|\lambda, I) z^2 \end{aligned} \tag{117}$$

with the derivative

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln \left(\frac{p(z|\lambda, I)}{m(z)} \right) - \lambda_1 z - \lambda_2 z^2 \\ &= 0, \end{aligned} \tag{118}$$

meaning

$$p(z|\lambda, I) = m(z)e^{-1-\lambda_0-\lambda_1 z-\lambda_2 z^2}. \tag{119}$$

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

$$\begin{aligned}\int dzp(z|\lambda, I) &= me^{-1-\lambda_0} \int dze^{-\lambda_1 z-\lambda_2 z^2} \\ &= me^{-1-\lambda_0} \sqrt{\frac{\pi}{\lambda_2}} e^{\frac{\lambda_1^2}{4\lambda_2}} \\ &= 1.\end{aligned} \tag{120}$$

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

$$\begin{aligned}p(z|\lambda, I) &= \frac{e^{-\lambda_1 x-\lambda_2 x^2}}{K} \\ &= \sqrt{\frac{\lambda_2}{\pi}} e^{-\frac{\lambda_1^2}{4\lambda_2}-\lambda_1 z-\lambda_2 z^2}.\end{aligned} \tag{121}$$

Now, imposing the mean constraint

$$\begin{aligned}\int dzzp(z|\lambda, I) &= \frac{\int dzze^{-\lambda_1 z-\lambda_2 z^2}}{K} \\ &= -\frac{\lambda_1}{2\lambda_2} \\ &= \mu.\end{aligned} \tag{122}$$

Hereby

$$\begin{aligned}p(z|\lambda, I) &= \sqrt{\frac{\lambda_2}{\pi}} e^{-\mu^2\lambda_2+2\mu\lambda_2 z-\lambda_2 z^2} \\ &= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{\mu-z}{\sigma}\right)^2},\end{aligned} \tag{123}$$

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 4.2.

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*

$$\begin{aligned}F = &- p(z|\lambda, I) \ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_0 p(z|\lambda, I) \\ &- \lambda_1 p(z|\lambda, I) \ln(z) - \lambda_2 p(z|\lambda, I) \ln(1 - z)\end{aligned} \tag{124}$$

with the derivative

$$\begin{aligned}\frac{\partial F}{\partial p(z|\lambda,I)} &= -1-\ln\left(\frac{p(z|\lambda,I)}{m(z)}\right)-\lambda_1\ln(z)-\lambda_2\ln(1-z)\\ &= 0,\end{aligned} \tag{125}$$

meaning

$$p(z|\lambda,I) = m(z)e^{-1-\lambda_0-\lambda_1\ln(z)-\lambda_2\ln(1-z)}. \tag{126}$$

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

$$\begin{aligned}\int dzp(z|\lambda,I) &= me^{-1-\lambda_0}\int dzz^{-\lambda_1}(1-z)^{-\lambda_2}\\ &= me^{-1-\lambda_0}\frac{\Gamma(1-\lambda_1)\Gamma(1-\lambda_2)}{\Gamma(2-\lambda_1-\lambda_2)}\\ &= 1.\end{aligned} \tag{127}$$

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

$$p(z|\alpha,\beta,I) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{128}$$

which is the beta distribution.

Example 4.3.

Consider a continuous random variable $Z \in [0,\infty)$ with known mean μ and known logarithmic mean ν. In this case

$$F = -p(z|\lambda,I)\ln\frac{p(z|\lambda,I)}{m(z)}-\lambda_0 p(z|\lambda,I)-\lambda_1 zp(z|\lambda,I)-\lambda_2\ln(z)p(z|\lambda,I), \tag{129}$$

and derivative

$$\begin{aligned}\frac{\partial F}{\partial p(z|\lambda,I)} &= -1-\ln\frac{p(z|\lambda,I)}{m(z)}-\lambda_0-\lambda_1 z-\lambda_2\ln z\\ &= 0\end{aligned} \tag{130}$$

meaning

$$\begin{aligned}p(z|\lambda,I) &= m(z)e^{-1-\lambda_0-\lambda_1 z-\lambda_2\ln z}\\ &= \tilde{m}(z)z^{-\lambda_2}e^{-\lambda_1 z}\end{aligned} \tag{131}$$

where $\tilde{m}(z) = m(z)e^{-1-\lambda_0}$. *Taking a uniform measure* ($m(z) = const$) *and imposing normalization*

$$\begin{aligned}\int_0^\infty dzp(z|\lambda, I) &= \tilde{m}\int_0^\infty dzz^{-\lambda_2}e^{-\lambda_1 z} \\ &= 1.\end{aligned} \tag{132}$$

The integral is recognized as the Gamma function

$$\int_0^\infty z^{\alpha-1}dze^{-\beta z} = \frac{\Gamma(\alpha)}{\beta^\alpha} \tag{133}$$

with $\alpha = 1 - \lambda_2$ *and* $\beta = \lambda_1$. *Substituting* $\tilde{m}$, α, β *back into Equation 131*

$$p(z|\lambda, I) = \frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}, \tag{134}$$

which is the Gamma distribution.

Example 4.4.

Consider a continuous random variable $Z \in [0, \infty)$ *with known mean* μ. *In this case*

$$F = -p(z|\lambda, I)\ln\frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \lambda_1 zp(z|\lambda, I), \tag{135}$$

and derivative

$$\frac{\partial F}{\partial p(z|\lambda, I)} = -1 - \ln\frac{p(z|\lambda, I)}{m(z)} - \lambda_0 - \lambda_1 z = 0. \tag{136}$$

Solving for $p(z|\lambda, I)$ *gives*

$$p(z|\lambda, I) = m(z)e^{-1-\lambda_0-\lambda_1 z}. \tag{137}$$

Taking $m(z) = const$ *and imposing normalization constraint*

$$\begin{aligned}\int_0^\infty dzp(z|\lambda, I) &= me^{-1-\lambda_0}\int_0^\infty dze^{-\lambda_1 z} \\ &= me^{-1-\lambda_0}\frac{1}{\lambda_1} \\ &= 1\end{aligned} \tag{138}$$

and the mean constraint

$$\begin{aligned}\int_0^\infty dzzp(z|\lambda, I) &= \int_0^\infty dzz\lambda_1 e^{-\lambda_1 z} \\ &= \frac{1}{\lambda_1} \\ &= \mu.\end{aligned} \tag{139}$$

Combining Equation 138, Equation 139 and Equation 137 yields

$$p(z|\lambda, I) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}, \tag{140}$$

which is the Gamma distribution.

Example 4.5.

Consider a random variable $Z \in \{0, 1\}$ with mean μ. In this case

$$\mathcal{L} = \sum_{z=0}^{1} F \tag{141}$$

F can be written

$$F = -p(z|\lambda, I) \ln \frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \lambda_1 z p(z|\lambda, I), \tag{142}$$

with the derivative

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_0 - \lambda_1 z \\ &= 0, \end{aligned} \tag{143}$$

meaning

$$p(z|\lambda, I) = m(z) e^{-1-\lambda_0-\lambda_1 z}. \tag{144}$$

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

$$\begin{aligned} \sum_{z=0}^{1} p(z) &= m e^{-1-\lambda_0} \left(1 + e^{-\lambda_1}\right) \\ &= 1 \end{aligned} \tag{145}$$

and mean constraint

$$\begin{aligned} \sum_{z=0}^{1} z p(z) &= m e^{-1-\lambda_0} e^{-\lambda_1} \\ &= \frac{1}{1 + e^{\lambda_1}} \\ &= \mu \end{aligned} \tag{146}$$

This means

$$\begin{aligned} p(0|\lambda, I) &= me^{-1-\lambda_0} \\ &= \frac{1}{1+e^{-\lambda_1}} \\ &= 1-\mu \end{aligned} \tag{147}$$

and

$$\begin{aligned} p(1|\lambda, I) &= me^{-1-\lambda_0-\lambda_1} \\ &= \mu, \end{aligned} \tag{148}$$

or

$$p(z|\lambda, I) = \mu^z(1-\mu)^{1-z}. \tag{149}$$

which is the Bernoulli distribution.

Example 4.6.

Consider $Z \in \{0, 1, \ldots, n\}$ representing the total number of successes in n independent Bernoulli trials with mean μ. Using the principle of maximum entropy, define

$$F = -p(z|\lambda, I) \ln \frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \lambda_1 z\, p(z|\lambda, I), \tag{150}$$

with the derivative

$$\begin{aligned} \frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln \frac{p(z|\lambda, I)}{m(z)} - \lambda_0 - \lambda_1 z \\ &= 0, \end{aligned} \tag{151}$$

which implies

$$p(z|\lambda, I) = m(z)e^{-\lambda_0-\lambda_1 z}. \tag{152}$$

Taking a uniform measure for the underlying sequences of Bernoulli trials, equivalent to $m(z) = \binom{n}{z}$, and imposing the normalization constraint

$$\begin{aligned} \sum_{z=0}^{n} p(z|\lambda, I) &= \sum_{z=0}^{n} \binom{n}{z} e^{-\lambda_0-\lambda_1 z} \\ &= 1, \end{aligned} \tag{153}$$

yields

$$e^{-\lambda_0} = (1+e^{-\lambda_1})^{-n}. \tag{154}$$

The mean constraint

$$\begin{aligned}\sum_{z=0}^{n} z\, p(z|\lambda, I) &= n\frac{e^{-\lambda_1}}{1+e^{-\lambda_1}} \\ &= n\mu\end{aligned} \tag{155}$$

gives

$$e^{-\lambda_1} = \frac{\mu}{1-\mu}. \tag{156}$$

Finally, substituting $e^{-\lambda_0}$ and $e^{-\lambda_1}$ into $p(z|\lambda, I)$ gives the maximum entropy distribution

$$p(z|\lambda, I) = \binom{n}{z}\mu^z(1-\mu)^{n-z}, \tag{157}$$

which is the Binomial distribution.

Example 4.7.

Consider a random variable $Z \in \{0, 1, 2, \dots\}$ with a known mean μ. In this case

$$\mathcal{L} = \sum_{z=0}^{\infty} F \tag{158}$$

and F can be written

$$F = -p(z|\lambda, I)\ln\frac{p(z|\lambda, I)}{m(z)} - \lambda_0 p(z|\lambda, I) - \lambda_1 z p(z|\lambda, I), \tag{159}$$

with the derivative

$$\begin{aligned}\frac{\partial F}{\partial p(z|\lambda, I)} &= -1 - \ln\left(\frac{p(z|\lambda, I)}{m(z)}\right) - \lambda_0 - \lambda_1 z \\ &= 0,\end{aligned} \tag{160}$$

meaning

$$p(z|\lambda, I) = m(z)e^{-1-\lambda_0-\lambda_1 z}. \tag{161}$$

For the Poisson distribution, the measure encodes the counting of configurations for each z and is given by $m(z) = 1/z!$. Imposing this measure and the normalization constraint

$$\begin{aligned}\sum_{z=0}^{\infty} p(z|\lambda, I) &= \sum_{z=0}^{\infty}\frac{e^{-1-\lambda_0-\lambda_1 z}}{z!} \\ &= e^{-1-\lambda_0}\sum_{z=0}^{\infty}\frac{e^{-\lambda_1 z}}{z!} \\ &= 1,\end{aligned} \tag{162}$$

Identifying the sum with the Taylor expansion

$$\sum_{z=0}^{\infty} \frac{e^{-\lambda_1 z}}{z!} = e^{e^{-\lambda}} \tag{163}$$

yields

$$e^{-1-\lambda_0} = e^{-e^{-\lambda_1}} \quad \Rightarrow \quad 1 + \lambda_0 = e^{-\lambda_1}. \tag{164}$$

Imposing the mean constraint

$$\begin{aligned} \sum_{z=0}^{\infty} z p(z|\lambda, I) &= e^{-1-\lambda_0} \sum_{z=1}^{\infty} \frac{z e^{-\lambda_1 z}}{z!} \\ &= e^{-1-\lambda_0} \sum_{z=1}^{\infty} \frac{e^{-\lambda_1 z}}{(z-1)!} \\ &= e^{-\lambda_1} e^{-1-\lambda_0} \sum_{y=0}^{\infty} \frac{e^{-\lambda_1 y}}{y!} \\ &= e^{-\lambda_1} \\ &= \mu \end{aligned} \tag{165}$$

where Equation 162 and $y = z - 1$ has been used. Combining Equation 161, Equation 164 and Equation 165 yield

$$p(z|\lambda, I) = \frac{\mu^z e^{-\mu}}{z!}. \tag{166}$$

which is the Poisson distribution.

CHAPTER 5

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, 16, 17]

$$X_i : \Omega \mapsto \Omega_{X_i}, \tag{167}$$

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

$$\mathbb{P}_{X_i} = \mathbb{P} \circ X_i^{-1} \tag{168}$$

is the image measure (see Definition 28) of $\mathbb{P}$ with respect to X_i.

Definition 38 (The Joint Probability Measure)**.** *The joint probability measure is the image measure*

$$\mathbb{P}_{X_1,\dots X_n} = \mathbb{P} \circ (X_1, \dots X_n)^{-1}. \tag{169}$$

on the measurable space

$$(\Omega_{X_1} \times \cdots \times \Omega_{X_n}, \mathcal{F}_{X_1} \otimes \cdots \otimes \mathcal{F}_{X_n}) \tag{170}$$

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 31) or continuous probability distributions (PDF, see Definition 32) associated to the joint probability measure. All probability distributions related to the random variables can be derived from the joint probability distribution via Theorem 3.

Definition 39 (Set of All Probability Measures)**.** *Let $\mathcal{P}$ be the set of all probability measures on $(\Omega_{X_{1:n}}, \mathcal{F}_{X_{1:n}})$.*

Definition 40 (Parametric Subset 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 the joint probability measure $\mathbb{P}_{X_{1:n}}$ *belongs to a parametric subset*

$$\mathcal{P}' = \{\, \mathbb{P}_{X_{1:n}}(w) \mid w \in \Omega_W \,\} \subseteq \mathcal{P}, \tag{171}$$

where Ω_W *is the parameter space.*

Definition 41 (Parameter Space)**.** *The parameter space* Ω_W *is the set of all values* w *that index the distributions* $\mathbb{P}_{X_1,\ldots,X_n}(w) \in \mathcal{P}'$*.*

Definition 42 (Identifiable Statistical Model)**.** *A statistical model is identifiable if the mapping*

$$w \in \Omega_W \mapsto \mathbb{P}_{X_1,\ldots,X_n}(w) \in \mathcal{P}' \tag{172}$$

is injective (i.e., distinct parameter values correspond to distinct distributions).

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* σ*-algebra (*$\mathcal{F}_W$*) and a probability measure (*$\mathbb{P}_W$*) that must be the result of another image measure (see Definition 28) with respect to the random variable* W*. This means*

$$W : \Omega \mapsto \Omega_W \tag{173}$$

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*

$$\mathbb{P}_W : \mathcal{F}_W \mapsto [0, 1], \tag{174}$$

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

$$\mathbb{P}_W = \mathbb{P} \circ W^{-1}. \tag{175}$$

Remark 2. *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.

5.1 INTERPRETATION OF PROBABILITY MEASURES

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 43 (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 [18]*

$$\mathbb{P}(E) \equiv \lim_{n \to \infty} \left(\frac{m}{n} \right) \tag{176}$$

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

Definition 44 (Sugeno Measure)**.** *Let* $(\Omega, \mathcal{F})$ *be a measurable space (Definition 22) and* $\text{Bel} : \mathcal{F} \to [0, 1]$ *a Sugeno measure iff [19]*

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

Definition 45 (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 44) [19, 20].*

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 46 (Frequentist Statistics)**.** *Frequentist statistics is a paradigm that adopts Axiom 4 and Definition 43 of probability.*

Definition 47 (Bayesian Statistics)**.** *Bayesian statistics is a paradigm that adopts Axiom 5 and definition Definition 45 of probability.*

Remark 3 (Frequentist vs. Bayesian Interpretation)**.** *In the Frequentist framework, parameters are fixed but unknown, and probability statements concern the variability of estimators across hypothetical repeated samples. For instance, a* 95% *confidence interval means that if the experiment were repeated many times, approximately* 95% *of the constructed intervals would contain the true parameter value.*

In the Bayesian framework, parameters are treated as random variables with a posterior distribution given the observed data. A 95% *credible interval therefore means that, conditional on the data and prior information, there is a* 95% *probability that the true parameter lies within the interval.*

Thus, in the Frequentist view, the interval varies across repeated experiments while the parameter remains fixed, whereas in the Bayesian view, the interval is fixed (given the data) and the parameter is uncertain.

Example 5.1.
Consider a Bayesian statistical model involving both a normal distribution with parameters μ, σ and a beta distribution with parameters a, b, then

$$W = \begin{pmatrix} W_\mu & W_\sigma & W_a & W_b \end{pmatrix}^T, \tag{177}$$

such that each individual parameter has an associated probability distribution.

Remark 4 (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 etc. 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.*

5.2 FRAMING OF STATISTICS

In this book, statistics is framed as a game against Nature, following conventions from decision theory[3]. In this game, there are two players, whose roles are formalized in Definition 48 and Definition 49.

Definition 48 (Robot)**.** *The Robot is the primary decision maker in the statistical game.*

Definition 49 (Nature)**.** *Nature is an unpredictable decision maker that can interfere with the Robot's outcomes. It models uncertainty in the decision-making process.*

Remark 5 (Statistical game setup)**.** *The game between the Robot and Nature is formalized by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a parameter space Ω_W, and a set of probability distributions $\mathcal{P}$ parameterized by $w \in \Omega_W$. The Robot and Nature each make a decision by choosing actions $u \in \Omega_U$ and $s \in \Omega_S$, respectively. The Robot receives a penalty from a cost function depending on both actions.*

Definition 50 (Cost Function)**.** *A cost function associates a numerical penalty depending on decision $u \in \Omega_U$ and $s \in \Omega_S$,*

$$C : \Omega_U \times \Omega_S \mapsto \mathbb{R}. \tag{178}$$

Given the observation $X = x$ as well as a set of past observations and matching actions of Nature

$$D = \{(X = x_i, S = s_i)\}_{i=1}^{n}, \tag{179}$$

the Robot's objective is to formulate a decision rule that minimize the expected cost associated with its decisions[2].

Definition 51 (Decision Rule)**.** *A decision rule is a function U that prescribes an action based on the current observation and past data. Formally, let $x \in \Omega_X$ be a new observation and $D \in (\Omega_X \times \Omega_S)^n$ denote the past observations and corresponding actions of Nature. Then a decision rule is a mapping*

$$U : \Omega_X \times (\Omega_X \times \Omega_S)^n \to \Omega_U, \tag{180}$$

where Ω_U is the action space of the Robot.

Example 5.2.

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

$$\Omega_U = \{\text{"}bring\ umbrella\text{"}, \text{"}don't\ bring\ umbrella\text{"}\}. \tag{181}$$

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

$$\Omega_S = \{\text{"}rain\text{"}, \text{"}no\ rain\text{"}\}, \tag{182}$$

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, $U(\text{weather forecast}, D)$ 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. Ω_X and Ω_S have associated σ-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 data D, the objective of the Robot is to minimize the expected cost associated with its decisions[2]

$$\begin{aligned}\mathbb{E}[C(U,S)|I] &= \int dDdxdsC(U(x,D),s)p(X=x,S=s,D|I) \\ &= \int d\tilde{D}dsC(U(\tilde{D}),s)p(S=s,\tilde{D}|I)\end{aligned} \tag{183}$$

where $\tilde{D} = \{D, X = x\}$, I denotes the background information (Definition 37) and the Robot aims to find the decision rule (Definition 51) which minimizes Equation 183, meaning

$$U^* = \arg\min_U \mathbb{E}[C(U,S)|I]. \tag{184}$$

From Theorem 10

$$\mathbb{E}[C(U,S)|I] = \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]]. \tag{185}$$

Using Equation 185 in Equation 184

$$\begin{aligned}U^* &= \arg\min_U \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]] \\ &= \arg\min_U \int d\tilde{D}p(\tilde{D}|I)\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I].\end{aligned} \tag{186}$$

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

$$\begin{aligned}U^*(\tilde{D}) &= \arg\min_{U(\tilde{D})} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] \\ &= \arg\min_{U(\tilde{D})} \int dsC(U(\tilde{D}),s)p(s|\tilde{D},I).\end{aligned} \tag{187}$$

Example 5.3.

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

$$\begin{aligned}\tilde{D} &= \{S = s_i\}_{i=1}^n \\ &\equiv D_s.\end{aligned} \tag{188}$$

In this case, the Robot is forced to model the decisions of Nature with a probability distribution with associated parameters without observations. From Equation 187 the optimal action for the Robot can be written

$$U^*(D_s) = \arg\min_{U(D_s)} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] \tag{189}$$

5.2.1 *Assigning a Cost Function*

The cost function (see definition 50) 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, Ω_U.

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*

$$C(U(\tilde{D}), s) \equiv |U(\tilde{D}) - s|. \tag{190}$$

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

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \int_{-\infty}^{\infty} ds|U(\tilde{D}) - s|p(s|\tilde{D}, I) \\ &= \int_{-\infty}^{U(\tilde{D})} ds(s - U(\tilde{D}))p(s|\tilde{D}, I) \\ &\quad + \int_{U(\tilde{D})}^{\infty} ds(U(\tilde{D}) - s)p(s|\tilde{D}, I)\end{aligned} \tag{191}$$

$$\begin{aligned}0 &= \frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}, I) + \int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D}, I) \\ &\quad + (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}, I) - \int_{U^*(\tilde{D})}^{\infty} dsp(s|\tilde{D}, I)\end{aligned} \tag{192}$$

$$\begin{aligned}\int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D},I) &= \int_{U^*(\tilde{D})}^{\infty} dsp(s|\tilde{D},I) \\ &= 1 - \int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D},I)\end{aligned} \tag{193}$$

$$\int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D},I) = \frac{1}{2} \tag{194}$$

which is the definition of the median.

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

$$C(U(\tilde{D}),s) \equiv (U(\tilde{D}) - s)^2. \tag{195}$$

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

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] &= \int ds(U(\tilde{D}) - s)^2 p(s|\tilde{D},I) \\ &\Downarrow \\ \frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(x)} &= 2U^*(\tilde{D}) - 2\int dssp(s|\tilde{D},I) \\ &= 0 \\ &\Downarrow \\ U^*(\tilde{D}) &= \int dssp(s|\tilde{D},I) \\ &= \mathbb{E}_{S|\tilde{D}}[S|\tilde{D},I]\end{aligned} \tag{196}$$

which is the definition of the expectation value.

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

$$C(U(\tilde{D}),s) \equiv 1 - \delta(U(\tilde{D}) - s). \tag{197}$$

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

$$\mathbb{E}_{S|\tilde{D}}[C((\tilde{D}),S)|\tilde{D},I] = 1 - \int ds\delta(U(\tilde{D}) - s)p(S = s|\tilde{D},I) \tag{198}$$

meaning

$$\begin{aligned}\frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} &= -\frac{\partial p(S = U(\tilde{D})|\tilde{D}, I)}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= 0\end{aligned} \tag{199}$$

which is the definition of the MAP.

Example 5.4.

The median decision rule is symmetric with respect to $z(\tilde{D}, s) \equiv U(\tilde{D}) - s$, meaning underestimation ($z < 0$) and overestimation ($z > 0$) is penalized equally. This decision rule can be generalized to favoring either scenario by adopting the cost function

$$C(U(\tilde{D}), s) = \alpha \cdot \text{swish}(U(\tilde{D}) - s, \beta) + (1 - \alpha) \cdot \text{swish}(s - U(\tilde{D}), \beta), \tag{200}$$

where

$$\text{swish}(z, \beta) = \frac{z}{1 + e^{-\beta z}}. \tag{201}$$

Taking $\alpha \ll 1$ means $z < 0$ will be penalized relatively more than $z > 0$. The expected cost is

$$\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] = \int dsp(S = s|\tilde{D}, I)\, C(U(\tilde{D}), s). \tag{202}$$

The derivative of the expected cost with respect to the decision rule can be approximated viz

$$\begin{aligned}\frac{dC}{dU} &= \frac{dC}{dz}\frac{dz}{dU} \\ &= \left(\frac{\alpha}{1 + e^{-\beta z}} - \frac{1 - \alpha}{1 + e^{\beta z}}\right. \\ &\qquad \left. + \frac{\alpha\beta e^{-\beta z} z}{(1 + e^{-\beta z})^2} + \frac{(1 - \alpha)\beta e^{\beta z} z}{(1 + e^{\beta z})^2}\right)\frac{dz}{dU} \\ &= \frac{\beta z e^{\beta z} - e^{\beta z} - 1}{(1 + e^{\beta z})^2} + \alpha + \mathcal{O}(\alpha^2) \\ &\approx \alpha - \frac{1}{(1 + e^{\beta z})^2}\end{aligned} \tag{203}$$

leading to the approximate expected cost

$$\begin{aligned}\frac{d\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I]}{dU(\tilde{D})} &\approx \int dsp(s|\tilde{D},I)\left(\alpha - \frac{1}{(1+e^{\beta z(\tilde{D},s)})^2}\right)\\ &= \alpha - \int dsp(s|\tilde{D},I)\frac{1}{(1+e^{\beta z(\tilde{D},s)})^2}.\\ &= 0\end{aligned} \tag{204}$$

For large β, the factor $\frac{1}{(1+e^{\beta(U(\tilde{D})-s)})^2}$ approaches the indicator $\mathbb{1}\{s > U(\tilde{D})\}$. Hence,

$$\int_{-\infty}^{\infty} dsp(s|\tilde{D},I)\frac{1}{(1+e^{\beta z(\tilde{D},s)})^2} \approx \int_{U(\tilde{D})}^{\infty} dsp(s|\tilde{D},I) \tag{205}$$

This means the optimal decision rule can be written viz

$$\alpha \approx \int_{U(\tilde{D})}^{\infty} dsp(s|\tilde{D},I). \tag{206}$$

The optimal decision $U^*(\tilde{D})$ is the α-quantile of the conditional distribution $p(S|\tilde{D},I)$. This rule is known as the quantile decision rule.

Discrete Action Space

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

$$\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] = \sum_{s\in\Omega_S} C(U(\tilde{D}),s)p(s|\tilde{D},I), \tag{207}$$

where the cost function is typically represented in matrix form viz

		S		
		$s^{(1)}$	$\dots$	$s^{(\dim(\Omega_S))}$
$U(\tilde{D})$	$u^{(1)}$	$C(u^{(1)},s^{(1)})$	$\dots$	$C(u^{(1)},s^{(\dim(\Omega_S))})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$u^{(\dim(\Omega_U))}$	$C(u^{(\dim(\Omega_U))},s^{(1)})$	$\dots$	$C(u^{(\dim(\Omega_U))},s^{(\dim(\Omega_S))})$

Note that the upper index represent realized values of s whereas a lower index represent datapoints.

Example 5.5.

With reference to Example 5.2, the possible states of Nature are $s^{(1)}$ = "rain" and $s^{(2)}$ = "no rain", whereas each observed outcome s_i in the dataset

$$D = \{(x_1, s_1), (x_2, s_2), (x_3, s_3)\} \tag{208}$$

takes a value in $\{s^{(1)}, s^{(2)}\}$. For instance, one possible dataset realization could be $s_1 = s^{(1)}$, $s_2 = s^{(1)}$, and $s_3 = s^{(2)}$.

Example 5.6.

Consider a binary classification problem with action space $\Omega_U = \{u^{(1)}, u^{(2)}\}$ and Nature's state space $\Omega_S = \{s^{(1)}, s^{(2)}\}$, where $u^{(1)}$ corresponds to predicting class $s^{(1)}$ and $u^{(2)}$ to predicting class $s^{(2)}$. Let

$$D = \{(x_i, s_i)\}_{i=1}^{n} \tag{209}$$

denote the training data, where $s_i \in \Omega_S$ are observed realizations of Nature's states. Let $U(x, D)$ be a classifier based on the probability $p(S = s|x, D, I)$. Define a threshold $k \in [0, 1]$ and the decision rule

$$U_k(x, D) = \begin{cases} u^{(1)}, & p(S = s^{(2)}|x, D, I) < k, \\ u^{(2)}, & p(S = s^{(2)}|x, D, I) \geq k. \end{cases} \tag{210}$$

For a fixed threshold k, classifier performance is summarized in the confusion matrix

		S	
		$s^{(1)}$	$s^{(2)}$
$U(x, D)$	$u^{(1)}$	*TP(k)*	*FP(k)*
	$u^{(2)}$	*FN(k)*	*TN(k)*

and standard performance measures are defined as

$$\mathrm{TPR(k)} = \frac{\mathrm{TP(k)}}{\mathrm{TP(k)} + \mathrm{FN(k)}}, \tag{211}$$

$$\mathrm{FPR(k)} = \frac{\mathrm{FP(k)}}{\mathrm{FP(k)} + \mathrm{TN(k)}}, \tag{212}$$

$$\mathrm{Accuracy(k)} = \frac{\mathrm{TP(k)} + \mathrm{TN(k)}}{\mathrm{TP(k)} + \mathrm{TN(k)} + \mathrm{FP(k)} + \mathrm{FN(k)}}. \tag{213}$$

Varying the threshold k over $[0, 1]$ defines a family of classifiers $U_k(x, D)$, which induces a set of points

$$\mathrm{ROC} = \{(\mathrm{FPR(k)}, \mathrm{TPR(k)}) : k \in [0, 1]\}. \tag{214}$$

The Area Under the ROC Curve (AUROC) is a threshold-independent measure. Let $X_{(s^{(1)})}$ and $X_{(s^{(2)})}$ denote independent draws from the class-conditional distributions $p(x|S = s^{(1)})$ and $p(x|S = s^{(2)})$, respectively. Then

$$\text{AUROC} = p(p(S = s^{(2)}|X_{(s^{(2)})}, D, I) > p(S = s^{(2)}|X_{(s^{(1)})}, D, I)|D, I), \quad (215)$$

i.e., the probability that the classifier assigns a higher score to a randomly chosen positive instance than to a randomly chosen negative instance. Equivalently,

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(u))\, du, \quad (216)$$

under regularity conditions ensuring FPR *is invertible. The Accuracy Ratio (AR), or normalized Gini coefficient, is defined from the AUROC as*

$$\text{AR} = 2 \cdot \text{AUROC} - 1. \quad (217)$$

and provide a measure rescaled to the interval $[-1, 1]$.

Example 5.7.

Consider a discrete action space with an observation $X = x$ and available data D ($\tilde{D} \equiv x, D$). 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

		S	
		$s^{(1)}$	$s^{(2)}$
$U(\tilde{D})$	$u^{(1)}$	0	λ_{12}
	$u^{(2)}$	λ_{21}	0

1. *Show that the decision u that minimizes the expected loss is equivalent to setting a probability threshold k and predicting $U(\tilde{D}) = u^{(1)}$ if $p(S = s^{(2)}|\tilde{D}, I) < k$ and $U(\tilde{D}) = u^{(2)}$ if $p(S = s^{(2)}|\tilde{D}, I) \geq k$. What is k as a function of λ_{12} and λ_{21}?*

   The conditional expected cost (Equation 207)

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \sum_s C(U(\tilde{D}), S = s)p(S = s|\tilde{D}, I) \\ &= C(U(\tilde{D}), S = s^{(1)})p(S = s^{(1)}|\tilde{D}, I) \\ &\quad + C(U(\tilde{D}), S = s^{(2)})p(S = s^{(2)}|\tilde{D}, I) \end{aligned} \quad (218)$$

For the different possible actions

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(u^{(1)},S)|\tilde{D},I] &= \lambda_{12}p(S=s^{(2)}|\tilde{D},I),\\ \mathbb{E}_{S|\tilde{D}}[C(u^{(2)},S)|\tilde{D},I] &= \lambda_{21}p(S=s^{(1)}|\tilde{D},I),\end{aligned} \tag{219}$$

$U(\tilde{D}) = u_1$ *iff*

$$\mathbb{E}_{S|\tilde{D}}[C(u^{(1)},S)|\tilde{D},I] < \mathbb{E}_{S|\tilde{D}}[C(u^{(1)},S)|\tilde{D},I]) \tag{220}$$

meaning

$$\begin{aligned}\lambda_{12}p(S=s^{(2)}|\tilde{D},I) &< \lambda_{21}p(S=s^{(1)}|\tilde{D},I)\\ &= \lambda_{21}(1-p(S=s^{(2)}|\tilde{D},I))\end{aligned} \tag{221}$$

meaning $U(\tilde{D}) = u_1$ *iff*

$$p(S=s^{(2)}|\tilde{D},I) < \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} = k \tag{222}$$

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

$k = \frac{1}{21} = \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} \Rightarrow \lambda_{12} = 9\lambda_{21}$ *yielding the loss matrix*

		S	
		$s^{(1)}$	$s^{(2)}$
$U(\tilde{D})$	$u^{(1)}$	0	$9\lambda_{21}$
	$u^{(2)}$	λ_{21}	o

You may set $\lambda_{21} = 1$ *since only the relative magnitude is important in relation to making a decision.*

Example 5.8.

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*

$$C(U(\tilde{D}),s) = \begin{cases} 0 & \text{if correct classification } (U(\tilde{D}) = s)\\ \lambda_r & \text{if reject option } (U(\tilde{D}) = \text{reject})\\ \lambda_s & \text{if wrong classification } (U(\tilde{D}) \neq s)\end{cases}. \tag{223}$$

1. *Show that the minimum cost is obtained if the robot decides on class* $U(\tilde{D})$ *if*

$$p(S = U(\tilde{D})|\tilde{D}, I) \geq p(S \neq U(\tilde{D})|\tilde{D}, I) \tag{224}$$

and if

$$p(S = U(\tilde{D})|\tilde{D}, I) \geq 1 - \frac{\lambda_r}{\lambda_s}. \tag{225}$$

The conditional expected cost if the robot does not pick the reject option, meaning $U(\tilde{D}) \in \Omega_U \setminus \text{reject}$

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \sum_s C(U(\tilde{D}), S = s)p(S = s|\tilde{D}, I) \\ &= \sum_{s \neq U(\tilde{D})} \lambda_s p(S = s|\tilde{D}, I) \\ &= \lambda_s(1 - p(S = U(\tilde{D})|\tilde{D}, I)) \end{aligned} \tag{226}$$

where for the second equality it has been used that the cost of a correct classification is 0*, so the case of* $S = U(\tilde{D})$ *does not enter the sum. For the third equality it has been used that summing over all but* $S = U(\tilde{D})$ *is equal to* $1 - p(S = U(\tilde{D})|\tilde{D}, I)$*. The larger* $p(S = U(\tilde{D})|\tilde{D}, I)$*, 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*

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(\text{reject}, S)|\tilde{D}, I] &= \lambda_r \sum_s p(S = s|\tilde{D}, I) \\ &= \lambda_r. \end{aligned} \tag{227}$$

Equation (226) *show picking* $\arg\max_{U(\tilde{D}) \in \Omega_U \setminus \text{reject}} p(S = U(\tilde{D})|\tilde{D}, I)$ *is the best option among classes* $U(\tilde{D}) \neq \text{reject}$*. To be the best option overall, it also needs to have lower cost than the reject option. Using Equation 226 and Equation 227 yields*

$$(1 - p(S = U(\tilde{D})|\tilde{D}, I))\lambda_s < \lambda_r \tag{228}$$

meaning

$$p(S = U(\tilde{D})|\tilde{D}, I) \geq 1 - \frac{\lambda_r}{\lambda_s}. \tag{229}$$

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

$$\frac{\lambda_r}{\lambda_s} = 0 \tag{230}$$

means rejection is rated as a successful classification – i.e. no cost associated – and this become the best option (rejection that is) unless

$$p(S = U(\tilde{D})|\tilde{D}, I) = 1, \tag{231}$$

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.

$$\frac{\lambda_r}{\lambda_s} = 1 \tag{232}$$

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.*

Remark 6 (Connection to Statistical Paradigms)**.** *So far in this chapter, there has been no reference to statistical paradigms (Bayesian or Frequentist). This is because all preceding material is valid under both the Bayesian (Definition 47) and Frequentist (Definition 46) paradigms. The difference between the two becomes apparent when considering the parameters of Nature's model.*

5.3 BAYESIAN STATISTICS

Bayesian statistics is based on Definition 47, which follows the definition of subjective probability (Definition 45) and the treating the parameters as realizations of a random variable (Axiom 5). The Bayesian framework originally come from the work of Bayes [21] and Laplace [22] with much of the modern discussions and formalism created later by Finetti [23] and Jeffreys [24] and Savage [25].

In the Bayesian paradigm, it is assumed that Natures decisions can be captured by a statistical model with parameters that are modeled as realizations of random variables. This means that the probability $p(S = s|X = x, D, I)$ in equation Equation 187 depend on the parameters $w_1, \ldots w_n$ of the statis-

tical 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

$$\begin{aligned} p(s|x, D, I) &= \int dw p(w, s|x, D, I) \\ &= \int dw p(s|w, x, D, I) p(w|x, D, I) \end{aligned} \tag{233}$$

Example 5.9.
Writing out the shorthand notation

$$p(W = w_1, \dots, W = w_n, S = s|X = x, D, I) \to p(w, s|x, D, I), \quad dw_1 \dots dw_n \to dw. \tag{234}$$

To evaluate $p(w|D, I)$ a combination of the chain rule (Theorem 1), Bayes' theorem (Theorem 2) and marginalization (Theorem 3) can be employed viz

$$\begin{aligned} p(w|x, D, I) &= p(w|D, I) \\ &= \frac{p(D_s|w, D_x, I) p(w|I)}{p(D_s|D_x, I)}, \end{aligned} \tag{235}$$

where $D_s = \{s_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ 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 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 (Definition 37). From Equation 235, 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.

5.3.1 *Bayesian Regression*

Regression involves the Robot building a model,

$$f : \Omega_W \times \Omega_X \mapsto \mathbb{R}, \tag{236}$$

with associated parameters $w \in \Omega_W$, that estimates Nature's actions $s \in \Omega_S = \mathbb{R}$ based on observed data $x \in \Omega_X$. Note that the output of f is $\mathbb{R}$ implying that S is assumed continuous. 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 from Definition 53 will be considered to review the subject. From Theorem 15 the best action for the Robot can be written

$$U^*(x, D) = \int dssp(s|x, D, I) \tag{237}$$

Assuming the actions of Nature follow a normal distribution with the function f as mean and an unknown precision, $\xi \in \Omega_W$

$$p(s|x, w, \xi, I) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(w,x)-s)^2}. \tag{238}$$

Using Equation 238 and marginalizing over ξ, w

$$\begin{aligned} p(s|x, D, I) &= \int dwd\xi p(s, w, \xi|x, D, I) \\ &= \int dwd\xi p(s|x, w, \xi, D, I)p(w, \xi|x, D, I) \\ &= \int dwd\xi p(s|x, w, \xi, I)p(w, \xi|D, I), \end{aligned} \tag{239}$$

where it has been used that $p(s|w, \xi, x, D, I) = p(s|w, \xi, x, I)$ since by definition f produce a $1-1$ map of the input x (Equation 238) and $p(w, \xi|x, D, I) = p(w, \xi|D, I)$ from Axiom 6. Using Equation 239 in Equation 237[1]

$$\begin{aligned} U^*(x, D) &= \int dwd\xi f(w, x)p(w, \xi|D, I), \\ &= \mathbb{E}[f|x, D, I] \end{aligned} \tag{240}$$

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

where it has been used that

$$\begin{aligned}\mathbb{E}[S|x,w,\xi,I] &= \int dssp(s|x,w,\xi,I) \\ &= f(w,x)\end{aligned} \tag{241}$$

according to Equation 238. Using Bayes theorem (Theorem 2)

$$p(w,\xi|D,I) = \frac{p(D_s|D_x,w,\xi,I)p(w,\xi|D_x,I)}{p(D_s|D_x,I)} \tag{242}$$

where from marginalization (Theorem 3)

$$p(D_s|D_x,I) = \int dwd\xi p(D_s|D_x,w,\xi,I)p(w,\xi|D_x,I). \tag{243}$$

Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written (using equation Equation 238)

$$p(D_s|D_x,w,\xi,I) = \left(\frac{\xi}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{\xi}{2}(f(w,x_i)-s_i)^2} \tag{244}$$

From the chain rule (see Theorem 1) and Theorem 6

$$p(w,\xi|D_x,I) = p(w|\xi,I)p(\xi|I). \tag{245}$$

Assuming the distributions of the w's are i) independent of ξ and ii) normally distributed[2] with zero mean and a precision described by a hyperparameter, λ.

$$\begin{aligned}p(w|\xi,I) &= p(w|I) \\ &= \int d\lambda p(w|\lambda,I)p(\lambda|I)\end{aligned} \tag{246}$$

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

$$p(w|\lambda,I)p(\lambda|I) = \prod_{q=1}^{\tilde{n}} \frac{\lambda_q^{\frac{n_q}{2}}}{(2\pi)^{\frac{n_q}{2}}} e^{-\frac{\lambda_q}{2}\sum_{l=1}^{n_q} w_l^2} \frac{\beta_q^{\alpha_q}}{\Gamma(\alpha_q)} \lambda_q^{\alpha_q-1} e^{-\beta_q\lambda_q} \tag{247}$$

where α_q, β_q are prior parameters (a part of the background information, Definition 37) and $\tilde{n}$ is the number of hyper parameters. In the completely general

2 The normally distributed prior is closely related to weight decay [26], a principle conventionally used in Frequentist statistics to avoid the issue of overfitting.

case $\tilde{n}$ would equal the number of parameters w, 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

$$\begin{aligned} p(\xi|I) &= \mathrm{Ga}(\xi|\tilde{\alpha}, \tilde{\beta}) \\ &= \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \xi^{\tilde{\alpha}-1} e^{-\tilde{\beta}\xi}. \end{aligned} \tag{248}$$

At this point equation Equation 237 is fully specified (the parameters $\alpha, \beta, \tilde{\alpha}, \tilde{\beta}$ and the functional form of $f(w, x)$ are assumed specified as part of the background information, Definition 37) and can be approximated by obtaining samples from $p(w, \xi, \lambda|D, I)$ via Hamiltonian Monte Carlo (HMC) [27–30] (see Appendix A for a review of HMC). The centerpiece in the HMC algorithm is the Hamiltonian defined viz [29, 30]

$$H \equiv \sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \ln[p(w, \xi, \lambda|D, I)] + const, \tag{249}$$

where

$$p(w, \xi|D, I) = \int d\lambda p(w, \xi, \lambda|D, I). \tag{250}$$

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 [31–33]). Using Equation 242-Equation 250 yields

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} + \frac{n}{2}[\ln(2\pi) - \ln(\xi)] + \frac{\xi}{2} \sum_{i=1}^{n} (f(w, x_i) - s_i)^2 \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln(\lambda_q) + \beta_q \lambda_q \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} w_l^2 \Bigg) \\ &+ \ln(\Gamma(\tilde{\alpha})) - \tilde{\alpha} \ln(\tilde{\beta}) + (1 - \tilde{\alpha}) \ln(\xi) + \tilde{\beta}\xi + const. \end{aligned} \tag{251}$$

Example 5.10.
Let $\xi \equiv e^{\zeta}$, such that $\zeta \in [-\infty, \infty]$ maps to $\xi \in [0, \infty]$ and ξ is ensured to be positive definite regardless of the value of ζ. Using the differential $d\xi = \xi d\zeta$ in Equation 240 means $p(\theta, \xi, \lambda|D, I)$ is multiplied with ξ. Hence, when taking $-\ln(p(\theta, \xi, \lambda|D, I))$ according to Equation 249, a $-\ln(\xi)$ is added to the Hamiltonian. In practice this means

$$(1 - \tilde{\alpha}) \ln(\xi) \in H \Rightarrow -\tilde{\alpha} \ln(\xi). \tag{252}$$

5.3.2 *Bayesian Classification*

Classification involves the Robot building a model,

$$f : \Omega_W \times \Omega_X \to \Delta^K, \tag{253}$$

with associated parameters $w \in \Omega_W$, that estimates Nature's actions $s \in \Omega_S = \{1, \ldots, K\}$ based on observed data $x \in \Omega_X$. Here

$$\Delta^K = \{p \in \mathbb{R}^K \mid p_s \geq 0, \sum_{s=1}^{K} p_s = 1\} \tag{254}$$

denotes the K-dimensional probability simplex, so that for each input $x \in \Omega_X$ the model output $f(w, x)$ is a probability vector representing the conditional distribution of the class label $s \in \Omega_S$. In particular, the probability of observing class s given x and parameters w is

$$p(S = s \mid x, w, I) = f_{S=s}(w, x), \tag{255}$$

where $f_{S=s}(w, x)$ denotes the s-th component of $f(w, x)$. By construction, these probabilities satisfy

$$\sum_{s \in \Omega_S} p(S = s \mid x, w, I) = 1. \tag{256}$$

In this case, the Robot's action space is equal to Natures action space, with the possible addition of a reject option, $\Omega_U = \Omega_S \cup \{\text{reject}\}$. To review 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 (Definition 54), with the addition of a reject option at cost λ. This means

$$C(U(\tilde{D}), s) = 1 - \delta_{U(\tilde{D}),s} + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}}. \tag{257}$$

The optimal decision rule for the robot can the be written (Equation 207)

$$\begin{aligned} U^*(\tilde{D}) &= \arg\min_{U(\tilde{D})} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] \\ &= \arg\min_{U(\tilde{D})} \left(\sum_{s \in \Omega_S} C(U(\tilde{D}), s) p(S = s|\tilde{D}, I) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right) \\ &= \arg\min_{U(\tilde{D})} \left(1 - p(S = U(\tilde{D})|\tilde{D}, I) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right). \end{aligned} \tag{258}$$

In absence of the reject option, the optimal decision rule is to pick the MAP, similar to Theorem 16. Using Equation 255 and marginalizing over w

$$\begin{aligned} p(S = U(\tilde{D})|\tilde{D}, I) &= \int dw p(S = U(\tilde{D}), w|\tilde{D}, I) \\ &= \int dw p(S = U(\tilde{D})|w, \tilde{D}, I) p(w|\tilde{D}, I) \\ &= \int dw p(S = U(\tilde{D})|x, w, I) p(w|D, I) \\ &= \int dw f_{S=U(\tilde{D})}(w, x) p(w|D, I) \\ &= \mathbb{E}[f_{S=U(\tilde{D})}(w, x)|D, I], \end{aligned} \tag{259}$$

where for the second to last equality it has been assumed that

$$p(S = U(\tilde{D})|w, \tilde{D}, I) = p(S = U(\tilde{D})|w, x, I) \tag{260}$$

since by definition f (see Equation 255) produce a $1 - 1$ map of the input x and $p(w|\tilde{D}, I) = p(w|D, I)$ from Axiom 6. From Bayes theorem

$$p(w|D, I) = \frac{p(D_s|D_x, w, I) p(w|D_x, I)}{p(D_s|D_x, I)}, \tag{261}$$

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

$$p(w|I) = \int d\lambda p(w|\lambda, I) p(\lambda|I). \tag{262}$$

where $p(w|\lambda, I)p(\lambda|I)$ is given by Equation 247. Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written [34]

$$\begin{aligned} p(D_s|D_x, w, I) &= \prod_{i=1}^{n} p(S = s_i|X = x_i, w, I) \\ &= \prod_{i=1}^{n} f_{S=s_i}(w, x_i) \end{aligned} . \tag{263}$$

At this point Equation 258 is fully specified and can be approximated by HMC similarly to the regression case (see Appendix A for a review of HMC). In this case, the model can be represented by the Hamiltonian

$$H \equiv \sum_q \sum_l \frac{p_l^2}{2m_l} - \ln(p(w, \lambda|D, I)) + const \tag{264}$$

where

$$p(w|D, I) = \int d\lambda p(w, \lambda|D, I). \tag{265}$$

Using Equation 259-Equation 263 in equation (264) yields the Hamiltonian

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \sum_{i=1}^{n} \ln(f_{S=s_i}(w, x_i)) + \text{const} \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln(\lambda_q) + \beta_q \lambda_q. \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} w_l^2 \Bigg) \end{aligned} \tag{266}$$

Sampling Equation 266 yields a set of coefficients which can be used to compute $\mathbb{E}[f_s(w, x)|D, I]$ which in turn (see Equation 258) can be used to compute $U^*(\tilde{D})$.

5.3.3 *Making Inference About the Model of Nature*

In some instances, the robot is interested in inference related to the model of Nature. The observation $x \in \Omega_X$ by definition does not have an associated known action of Nature and thus by Axiom 6 is disregarded in this context. From Equation 187

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{S|D}[C(U(D), S)|D, I] \tag{267}$$

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

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 267

$$\begin{aligned}\mathbb{E}[C(u_1, S)|D, I] &= \sum_{s=s_1,s_2} C(u_1, s)p(S = s|D, I),\\ \mathbb{E}[C(u_2, S)|D, I] &= \sum_{s=s_1,s_2} C(u_2, s)p(S = s|D, I),\end{aligned} \tag{268}$$

where in this case $u_i = s_i \quad \forall (u_i, s_i) \in \Omega_U \times \Omega_S$ 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

$$\frac{p(s_1|D, I)}{p(s_2|D, I)} > \frac{C(u_1, s_2) - C(u_2, s_2)}{C(u_2, s_1) - C(u_1, s_1)}. \tag{269}$$

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

$$\begin{aligned}\text{posterior ratio} &= \frac{p(s_1|D, I)}{p(s_2|D, I)}\\ &= \frac{p(D_s|s_1, D_x, I)p(s_1|I)}{p(D_s|s_2, D_x, I)p(s_2|I)},\end{aligned} \tag{270}$$

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, $p(s_1|I) = p(s_2|I)$

$$\text{posterior ratio} = \frac{p(D_s|s_1, D_x, I)}{p(D_s|s_2, D_x, I)}. \tag{271}$$

$p(D_s|s_1, D_x, I)$ 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 271 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.*

Bayesian Parameter Estimation

Let $w \in \Omega_W$ represent a parameter with the associated random variable W. 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 (Equation 187)

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{W|D}[C(U(D), W)|D, I], \tag{272}$$

with

$$\mathbb{E}_{W|D}[C(U(D), W)|D, I] = \int dw C(U(D), w) p(w|D, I). \tag{273}$$

At this point, the Robot can select a cost function like in Section 5.2.1 and proceed by expanding $p(w|D, I)$ similarly to Equation 235. Picking the quadratic cost (Definition 53) yields

$$U^*(D) = \mathbb{E}_{W|D}[W|D, I] \tag{274}$$

$p(w|D, I)$ in Equation 274 can be expanded as shown in Equation 235.

Example 5.11.

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

$$p(w_B > w_A|D, I) = \int_0^1 \int_{w_A}^1 p(w_A, w_B|D, I) dw_A dw_B. \tag{275}$$

Assuming the costumer sets are independent

$$\begin{aligned} p(w_A, w_B|D, I) &= p(w_B|w_A, D, I) p(w_A|D, I) \\ &= p(w_B|D_A, I) p(w_A|D_A, I), \end{aligned} \tag{276}$$

with

$$p(w_i|D_i, I) = \frac{p(D_i|w_i, I)p(w_i|I)}{p(D_i|I)}. \tag{277}$$

Assuming a beta prior and a binomial likelihood yields (since the binomial and beta distributions are conjugate)

$$p(w_i|D_i, I) = \frac{w_i^{\alpha_i - 1}(1 - w_i)^{\beta_i - 1}}{B(\alpha_i, \beta_i)}, \tag{278}$$

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 275 yields*

$$p(w_B > w_A|D, I) = \sum_{j=0}^{\alpha_B - 1} \frac{B(\alpha_A + j, \beta_A + \beta_B)}{(\beta_B + j)B(1 + j, \beta_B)B(\alpha_A, \beta_A)}. \tag{279}$$

5.4 FREQUENTIST STATISTICS

Frequentist statistics is based on Definition 46, which follows the definition of objective probability (Definition 43) 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 [35] and Fisher [36], who laid the groundwork for much of its methodology. Subsequent developments by Wald [37], Neyman [38], and Lehmann [39] further refined its theories and techniques.

In the Frequentist paradigm, it is assumed that Nature's actions are generated by a model with parameters $w \in \Omega_W$, which are unknown but fixed. In this setting, the optimal decision rule can be expressed as

$$U^*(x, w). \tag{280}$$

Thus, all quantities in Section 5.2 become conditioned on w. Since w is not known to the Robot, the central task becomes to estimate w from past data D.

This gives rise to a nested decision problem with two levels:

i) Parameter estimation: use past data D to construct an estimator $\hat{w}(D)$ of the fixed but unknown parameter w.

ii) Prediction/decision: given a new observation x and the parameter estimate $\hat{w}(D)$, apply the decision rule U to determine an action.

To avoid notational ambiguity, a distinction is made between the decision rule used for prediction, denoted U, and the decision rule used for parameter estimation, denoted $\hat{w}$. The practical decision rule for a new observation $x \in \Omega_X$ therefore takes the form

$$U^*(x, \hat{w}^*(D)), \tag{281}$$

where $\hat{w}^*(D)$ denotes the optimal parameter decision rule, obtained from past data D, and the final action is determined by minimizing the expected cost as specified in Section 5.2.

5.4.1 *Frequentist Regression*

In the Frequentist paradigm, regression involves the Robot constructing a model,

$$f : \Omega_W \times \Omega_X \mapsto \mathbb{R}, \tag{282}$$

parameterized by $w \in \Omega_W$, to approximate Nature's actions $s \in \Omega_S$ based on observed data $x \in \Omega_X$. As in Bayesian regression (Section 5.3.1), the output of f is real-valued, so that S is assumed continuous. The model f serves as the Robot's surrogate for Nature's mechanism, providing predictions of Nature's action given observed data $x \in \Omega_X$.

Suppose Nature's action $s \in \Omega_S$ given observed data $x \in \Omega_X$ is distributed according to a normal distribution with mean $f(w, x)$ and precision $\xi \in \Omega_\Xi$,

$$p(s|x, w, \xi, I) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(w,x)-s)^2} \tag{283}$$

where I denotes the background information (Definition 37). Here, (w, ξ) are fixed but unknown parameters. Under the quadratic cost function from Definition 53, the optimal decision rule is the conditional expectation of S given (x, w, ξ) (Theorem 15),

$$\begin{aligned} U^*(x, \hat{w}^*(D), \hat{\xi}^*(D)) &= \mathbb{E}[S|x, \hat{w}^*(D), \hat{\xi}^*(D), I] \\ &= \int s p(s|x, \hat{w}^*(D), \hat{\xi}^*(D), I) ds \\ &= f(\hat{w}^*(D), x). \end{aligned} \tag{284}$$

Equation 284 represents the Frequentist optimal decision rule conditional on the parameter estimate, whereas Equation 240 represents the Bayesian optimal decision rule, which averages over the posterior distribution of the model parameters (and latent variables) given the data. From equation Equation 284 it is clear that in Frequentist statistics, regression is reframed as parameter estimation.

5.4.2 *Frequentist Classification*

In the Frequentist paradigm, classification involves the Robot constructing a model

$$f : \Omega_W \times \Omega_X \mapsto \Delta^K, \tag{285}$$

parameterized by $w \in \Omega_W$, where Δ^K is the K-dimensional probability simplex and $\Omega_S \in \{1, \ldots, K\}$ represents Nature's discrete action (class label). The model predicts the conditional probability of each class given the input $x \in \Omega_X$

$$p(S = s|x, w, I) = f_s(w, x), \quad s \in \{1, \ldots, K\}, \tag{286}$$

with

$$\sum_{s \in \Omega_S} p(S = s|x, w, I) = 1. \tag{287}$$

The Robot's action space is typically equal to Nature's action space, $\Omega_U = \Omega_S$, possibly with the addition of a reject option at cost λ. Using the 0–1 cost function with optional reject,

$$C(U(x, w), s) = 1 - \delta_{U(x,w),s} + (\lambda - 1)\delta_{U(x,w),\text{reject}}. \tag{288}$$

Let $\hat{w}^*(D)$ denote the optimal Frequentist estimator of the model parameters obtained from past data D. The optimal decision rule for a new observation x is

$$\begin{aligned} U^*(x, \hat{w}^*(D)) &= \arg\min_{u \in \Omega_U} \mathbb{E}[C(u, S) \mid x, \hat{w}(D), I] \\ &= \arg\min_{u \in \Omega_U} \Big(1 - f_u(\hat{w}^*(D), x) + (\lambda - 1)\delta_{u,\text{reject}}\Big). \end{aligned} \tag{289}$$

From equation Equation 289 it is clear that in Frequentist statistics, classification is reframed as parameter estimation.

5.4.3 *Frequentist Parameter Estimation*

As shown in Section 5.4.1 and Section 5.4.2, both regression and classification in the Frequentist paradigm can be reframed as problems of parameter estimation. This makes parameter estimation the central focus of Frequentist statistics. Unlike in Bayesian statistics, where parameters are intermediate quantities to be marginalized over, in the Frequentist framework the parameters are fixed but unknown, and their determination carries substantive interpretational and practical importance. Estimators of these parameters serve as decision rules that summarize past observations into actionable predictions.

Definition 56 (Sampling distribution)**.** *Let D denote the observed dataset and let $\hat{w}(D)$ be a decision rule (estimator) for the fixed-but-unknown parameter $w \in \Omega_W$. The sampling distribution of $\hat{w}$ is the probability distribution of the random variable $\hat{w}(D)$ induced by repeated sampling of D from the data-generating mechanism $p(D \mid w, I)$.*

Remark 7 (Bayesian versus Frequentist perspective)**.** *The sampling distribution of an estimator $\hat{w}(D)$ is central to the Frequentist paradigm, since all uncertainty arises from the randomness of the data $D \sim p(D \mid w, I)$ while the parameter w is treated as a fixed but unknown constant. In Bayesian statistics, by contrast, uncertainty about w is represented by a posterior distribution $p(w \mid D, I)$ after observing data. Both approaches yield distributions over possible parameter values or estimates, but their conceptual origin differs: in the Frequentist case, the distribution is over repeated samples of data, whereas in the Bayesian case, the distribution is over the parameter itself given the observed data.*

Example 5.12.

In practice, the true sampling distribution of an estimator $\hat{w}(D)$ is rarely available in closed form. The bootstrap *provides an approximation technique based solely on the observed dataset. Let $D = \{(x_i, s_i)\}_{i=1}^{n}$ be the dataset. A bootstrap sample D^* is constructed by sampling n observations with replacement from D. Repeating this procedure B times yields bootstrap replicates $\hat{w}(D^{*1}), \ldots, \hat{w}(D^{*B})$, whose empirical distribution approximates the sampling distribution of $\hat{w}(D)$.*

Common quantities derived from the bootstrap include:

- *The bootstrap estimate of variance:*

$$\widehat{\mathrm{Var}}_{\mathrm{boot}}[\hat{w}] = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{w}(D^{*b}) - \overline{\hat{w}}^{*} \right)^2, \tag{290}$$

where $\overline{\hat{w}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{w}(D^{*b})$.

- *The bootstrap confidence interval, constructed from quantiles of the bootstrap distribution of* $\hat{w}$.

Definition 57 (Fisher Information)**.** *Take* $D_s = \{s_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ *and let* $w \in \Omega_W$ *be an unknown parameter of the model. Let* $p(D_s|D_x, w, I)$ *denote the likelihood of observing Natures actions* D_s *given observed data* D_x *and* w*. The Fisher information is defined as*

$$\begin{aligned}\mathcal{I}(w) &\equiv \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\right)^2\middle| D_x, w, I\right] \\ &= \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\middle| D_x, w, I\right].\end{aligned} \tag{291}$$

Proof. In general

$$\mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p\right)^2\right] = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p\right] + \left(\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right]\right)^2. \tag{292}$$

Now

$$\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right] = \int dD_s\left(\frac{\partial}{\partial w}\ln p\right)p \tag{293}$$

$$= \int dD_s \frac{\partial}{\partial w}p \tag{294}$$

$$= \frac{\partial}{\partial w}\int dD_s p \tag{295}$$

$$= 0, \tag{296}$$

since $\int dD_s\, p(D_s \mid D_x, w, I) = 1$. Therefore

$$\mathcal{I}(w) = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_s \mid D_x, w, I)\middle| D_x, w, I\right]. \tag{297}$$

□

Theorem 17 (Fisher Information for Independent Observations)**.** *Take* $D_s = \{s_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ *and let* $w \in \Omega_W$ *be a parameter of the model. Assume the likelihood factorizes as*

$$p(D_s|D_x, w, I) = \prod_{i=1}^{n} p(s_i|x_i, w, I). \tag{298}$$

Then, the Fisher information of the full dataset is

$$\mathcal{I}(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\right)^2\middle|D_x,w,I\right] = \sum_{i=1}^{n}\mathcal{I}_i(w), \tag{299}$$

where $\mathcal{I}_i(w)$ is the Fisher information of the i-th observation:

$$\mathcal{I}_i(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(s_i|x_i,w,I)\right)^2\middle|x_i,w,I\right]. \tag{300}$$

Definition 58 (Maximum Likelihood Estimator (MLE) Decision Rule)**.** *Take $D_s = \{s_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ and let $w \in \Omega_W$ be a fixed but unknown parameter. The Maximum Likelihood Estimator (MLE) decision rule $\hat{w}_{\text{MLE}}$ is the value of w that maximizes the likelihood of observing D_s given D_x*

$$\hat{w}_{\text{MLE}}(D) \equiv \arg\max_{w\in\Omega_W} p(D_s|D_x,w,I). \tag{301}$$

Theorem 18 (Asymptotic Sampling Distribution of the MLE)**.** *Let $\hat{w}_{\text{MLE}}(D)$ denote the Maximum Likelihood Estimator (MLE) of the fixed-but-unknown parameter w. Under standard regularity conditions, the sampling distribution of $\hat{w}_{\text{MLE}}$ satisfies*

$$\sqrt{n}\left(\hat{w}_{\text{MLE}} - w\right) \xrightarrow{d} \text{Norm}\left(0, \mathcal{I}(w)^{-1}\right), \tag{302}$$

where $\mathcal{I}(w)$ is the Fisher information matrix evaluated at w and $\xrightarrow{d}$ denotes convergence in distribution as $n \to \infty$. That is, the sampling distribution of the MLE becomes approximately normal, centered at the true parameter w with variance given by the inverse Fisher information.

Definition 59 (Minimax Decision Rule)**.** *A decision rule $\hat{w}'$ is said to be minimax if it minimize the maximum expected cost, meaning (Equation 183)*

$$\begin{aligned}\hat{w}' &\equiv \inf_{\hat{w}}\sup_{w\in\Omega_W}\mathbb{E}[C(\hat{w},w)|w,I] \\ &= \inf_{\hat{w}}\sup_{w\in\Omega_W}\int dD C(\hat{w}(D),w)p(D|w,I).\end{aligned} \tag{303}$$

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

$$\begin{aligned}\mathbb{E}[C(\hat{w},w)|w,I] &= \mathbb{E}[(\hat{w}-w)^2|w,I] \\ &= \mathbb{E}[(\hat{w}-\mathbb{E}[\hat{w}|I])^2|w,I] + (w-\mathbb{E}[\hat{w}|I])^2 \\ &= \text{Var}[\hat{w}|w,I] + \text{Bias}[\hat{w}|w,I]^2\end{aligned} \tag{304}$$

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

$$\text{Bias}[\hat{w}|w, I] \equiv w - \mathbb{E}[\hat{w}|I]. \tag{305}$$

If $\mathbb{E}[C(\hat{w}, w)|w, I] \to 0$ as $n \to \infty$, then $\hat{w}$ is a weakly consistent estimator of w, i.e., $\hat{w} \xrightarrow{p} w$. 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) $\hat{w}_{MLE}$ is approximately minimax for the quadratic cost function (Definition 53), meaning it approximately minimizes the maximum expected cost.*

Proof. From theorem Theorem 19

$$\mathbb{E}[(\hat{w} - w)^2|w, I] = \text{Var}[\hat{w}|w, I] + \text{Bias}[\hat{w}|w, I]^2. \tag{306}$$

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

$$\begin{aligned} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w, I] &\approx \text{Var}[\hat{w}_{\text{MLE}}|w, I] \\ &\approx \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \end{aligned} \tag{307}$$

where Theorem 18 was used for the second line. The Cramer-Rao lower bound [40] for variance states that

$$\text{Var}[\hat{w}|w, I] \geq \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \tag{308}$$

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

$$\sup_{w \in \Omega_W} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w, I] \approx \inf_{\hat{w}} \sup_{w \in \Omega_W} \mathbb{E}[(\hat{w} - w)^2|w, I], \tag{309}$$

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

Example 5.13.
The bias-variance decomposition (Theorem 19) is a concept relevant to Frequentist statistics, where a single point estimate of the parameters is used. This decomposition illustrates the tradeoff between underfitting and overfitting: high bias corresponds to underfitting, while high variance corresponds to overfitting.

In Bayesian statistics, predictions are obtained by integrating over the posterior distribution of parameters, rather than relying on a single point estimate. This integration inherently regularizes the model, mitigating overfitting and underfitting.

Example 5.14.
Take $D_s = \{S = s_i\}_{i=1}^n$ with $S \sim \mathrm{Ber}(w)$, and let $w \in [0,1]$ be the unknown parameter. Determine the quadratic cost of three different decision rules for estimating w: the arithmetic sample mean, the constant 0.5, and the first observation s_1.

- *Arithmetic mean:*

$$\hat{w}(D_s) = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{310}$$

with

$$\begin{aligned}
\mathbb{E}[\hat{w}(D_s)|w,I] &= \int dD_s \hat{w}(D_s) p(D_s|w,I) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[S|w,I] \\
&= w, \\
\mathrm{Var}[\hat{w}(D_s)|w,I] &= \int dD_s (\hat{w}(D_s) - \mathbb{E}[\hat{w}(D_s)|w,I])^2 p(D_s|w,I) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[S|w,I] \\
&= \frac{w(1-w)}{n}, \\
\mathbb{E}[(\hat{w}(D_s) - w)^2|w,I] &= \mathrm{Var}[\hat{w}(D_s)|w,I] + (\mathbb{E}[\hat{w}(D_s)|w,I] - w)^2 \\
&= \frac{w(1-w)}{n}.
\end{aligned} \tag{311}$$

- *Constant estimate:*

$$\hat{w} = 0.5 \tag{312}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}|w, I] &= 0.5, \\ \mathrm{Var}[\hat{w}|w, I] &= 0, \\ \mathbb{E}[(\hat{w} - w)^2|w, I] &= (0.5 - w)^2. \end{aligned} \tag{313}$$

- *First observation:*

$$\hat{w}(D_s) = s_1 \tag{314}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}(D_s)|w, I] &= \mathbb{E}[S|w, I] \\ &= w, \\ \mathrm{Var}[\hat{w}(D_s)|w, I] &= \mathrm{Var}[S|w, I] + (\mathbb{E}[\hat{w}(D_s)|w, I] - w)^2 \\ &= w(1 - w), \\ \mathbb{E}[(\hat{w}(D_s) - w)^2|w, I] &= w(1 - w). \end{aligned} \tag{315}$$

The arithmetic mean minimizes the quadratic cost over the entire range of w, while the constant 0.5 performs better for specific values of w. The cost of using s_1 is independent of n, making it less favorable as the sample size increases.

Example 5.15.

Take $D_s = \{S = s_i\}_{i=1}^n$ with $S \sim \mathrm{Ber}(w)$, and let $w \in [0, 1]$ be the unknown parameter. Determine the maximum likelihood estimate of w.

In this case

$$\begin{aligned} p(D_s|D_x, w, I) &= p(D_s|w, I) \\ &= \prod_{i=1}^{n} w^{s_i}(1 - w)^{1 - s_i}. \end{aligned} \tag{316}$$

Let $l(w) \equiv \ln p(D_s|D_x, w, I)$, then

$$\begin{aligned} \operatorname*{argmax}_w l(w) &= \operatorname*{argmax}_w p(D_s|w, I) \\ &= \operatorname*{argmax}_w \ln\left(\prod_{i=1}^{n} w^{s_i}(1 - w)^{1 - s_i}\right) \\ &= \operatorname*{argmax}_w \left[\ln w \sum_{i=1}^{n} s_i + \ln(1 - w) \sum_{i=1}^{n} (1 - s_i)\right] \end{aligned} \tag{317}$$

Now

$$\frac{d}{dw}l(w) = \frac{\sum_{i=1}^{n} s_i}{w} - \frac{n - \sum_{i=1}^{n} s_i}{1 - w} \tag{318}$$

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

$$\hat{w}_{MLE}(D_s) = \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{319}$$

Example 5.16.

Take $D_s = \{S = s_i\}_{i=1}^{n}$ with $S \sim \text{Exp}(w)$, and let $w > 0$ be the unknown parameter. Determine the maximum likelihood estimate of w.

In this case

$$\begin{aligned} p(D_s|D_x, w, I) &= p(D_s|w, I) \\ &= \prod_{i=1}^{n} we^{-ws_i}. \end{aligned} \tag{320}$$

Let $l(w) \equiv \ln p(D_s|D_x, w, I)$, then

$$\frac{d}{dw}l(w) = \frac{n}{w} - \sum_{i=1}^{n} s_i \tag{321}$$

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

$$\hat{w}_{MLE}(D_s) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} s_i}. \tag{322}$$

APPENDIX A

Hamiltonian Monte Carlo

This appendix is taken from Petersen [41]. The Hamiltonian Monte Carlo Algorithm (HMC algorithm) is a Markov Chain Monte Carlo (MCMC) algorithm used to evaluate integrals on the form

$$\begin{aligned}\mathbb{E}[f] &= \int f(\theta)g(\theta)d\theta \\ &\approx \frac{1}{N}\sum_{j\in g} f(\theta_j),\end{aligned} \tag{323}$$

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, θ, 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 θ and the transition probability $p(\theta'|\theta)$, then [29]

$$g(\theta') = \int p(\theta'|\theta)g(\theta)d\theta. \tag{324}$$

Equation 324 is implied by the stronger condition of detailed balance, defined viz

$$p(\theta'|\theta)g(\theta) = p(\theta|\theta')g(\theta'). \tag{325}$$

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 [42, 43]. 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 [28] 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 [30] 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

$$g = e^{\frac{F-H}{k_B T}} \Rightarrow H = F - k_B T \ln[g], \tag{326}$$

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 [44] $Z \sim \mathcal{N}(0, \sqrt{m_i})$ is taken yielding the Hamiltonian

$$H = -k_B T \left[\ln[g] - \sum_i \frac{p_i^2}{2m_i} \right] + const, \tag{327}$$

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 [30], however, the temperature can be used to manipulate the range of states which can be accepted e.g. via simulated annealing [45]. Here $T = k_b^{-1}$ will be adopted in accordance with [29, 30] and as such

$$H = \sum_i \frac{p_i^2}{2m_i} - \ln[g]. \tag{328}$$

The dynamics in parameter space are determined by Hamiltons equations

$$\dot{\theta}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial \theta_i}, \tag{329}$$

with θ_i denoting the different variables (coefficients). In order to implement Hamiltons equations, they are discretized via the leap frog method [29, 30] viz

$$\begin{aligned} p_i\left(t + \frac{\epsilon}{2}\right) &= p_i(t) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t), p_i(t))}{\partial \theta_i}, \\ \theta_i(t+\epsilon) &= \theta_i(t) + \frac{\epsilon}{m_i} p_i\left(t + \frac{\epsilon}{2}\right), \\ p_i(t+\epsilon) &= p_i\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t+\frac{\epsilon}{2}), p_i(t+\frac{\epsilon}{2}))}{\partial \theta_i}, \end{aligned} \tag{330}$$

with ϵ 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 ϵ 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 ϵ, 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

$$\mathbb{P}(\text{transition}) = \min\left[1, e^{H_{initial} - H_{final}}\right]. \tag{331}$$

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 ϵ 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 $\mathcal{O}(\epsilon^2)$ [29] and so the error can be controlled by ensuring that $\epsilon \ll 1$. A too small value of ϵ 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 ϵ can be derived from a one dimensional Gaussian Hamiltonian

$$H = \frac{q^2}{2\sigma^2} + \frac{p^2}{2}. \tag{332}$$

The leap frog step for this system is a linear map from $t \to t + \epsilon$. The mapping can be written

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \begin{bmatrix} 1 - \frac{\epsilon^2}{2\sigma^2} & \epsilon \\ \epsilon(\frac{1}{4}\epsilon^2\sigma^{-4} - \sigma^{-2}) & 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \tag{333}$$

The eigenvalues of the coefficient matrix represent the powers of the exponentials that are the solutions to the differential equation. They are given by

$$\text{Eigenvalues} = 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \pm \epsilon\sigma^{-1}\sqrt{\frac{1}{4}\epsilon^2\sigma^{-2} - 1}. \tag{334}$$

In order for the solutions to be bounded, the eigenvalues must be imaginary, meaning that

$$\epsilon < 2\sigma. \tag{335}$$

In higher dimensions a rule of thumb is to take $\epsilon \lesssim 2\sigma_x$, where σ_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 [44] a stable solution with $\frac{1}{2}p^T\Sigma^{-1}p$ as the kinetic term in the Hamiltonian require

$$\epsilon_i < 2\lambda_i^{-\frac{1}{2}}, \tag{336}$$

for each eigenvalue λ_i of the matrix

$$M_{ij} = (\Sigma^{-1})_{ij}\frac{\partial^2 H}{\partial q_i \partial q_j}, \tag{337}$$

which means that in the case of $\Sigma^{-1} = diag(m_i^{-1})$;

$$\epsilon_i < 2\sqrt{\frac{m_i}{\frac{\partial^2 H}{\partial q_i^2}}}. \tag{338}$$

Setting ϵ according to Equation 336 can however introduce issues for hierarchical models (models including hyper parameters) since the reversibility property of Hamiltonian dynamics is broken if ϵ depend on any parameters. This issue can be alleviated by using the MH algorithm on a subgroup of parameters [29, 30] (which are then allowed in the expression for ϵ) that is to be included in ϵ. However, unless the MH algorithm is used for all parameters, some degree of approximation is required.

Algorithm 1 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_{lower} and L_{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

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] E. T. Jaynes. "Probability Theory - The Logic of Science."

[9] E. T. Jaynes. "Prior Probabilities." In: *IEEE Transactions on Systems Science and Cybernetics* SSC-4 (1968), pp. 227–241.

[10] 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.

[11] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons, 1971.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Peter Orbanz. *Functional Conjugacy in Parametric Bayesian Models*. Technical Report. University of Cambridge, 2009.

[17] Daniel V. Tausk. *A Basic Introduction to Probability and Statistics for Mathematicians*. Date: January 24th, 2023. 2023.

[18] Edward E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, 1978, p. 25.

[19] Glenn Shafer. "BELIEF FUNCTIONS AND POSSIBILITY MEASURES." English (US). In: *Anal of Fuzzy Inf.* CRC Press Inc, 1987, pp. 51–84. ISBN: 0849362962.

[20] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer, 2009. DOI: 10.1007/978-0-387-92407-6.

[21] 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.

[22] Pierre-Simon Laplace. *ThÃľorie analytique des probabilitÃľs*. Paris: Courcier, 1812. URL: http://gallica.bnf.fr/ark:/12148/bpt6k88764q.

[23] 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.

[24] 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.

[25] L. Savage. *The Foundations of Statistics*. New York: Wiley, 1954.

[26] 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.

[27] J. M Hammersleyand and D. C. Handscomb. *Monte Carlo Methods*. London, Methuen., 1964.

[28] 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.

[29] Radford M. Neal. Berlin, Heidelberg: Springer-Verlag, 1996. ISBN: 0387947248.

[30] Radford M. Neal. "MCMC using Hamiltonian dynamics." In: (2012). cite arxiv:1206.1901. URL: http://arxiv.org/abs/1206.1901.

[31] 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/.

[32] 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.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618.

[34] 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.

[35] 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.

[36] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[37] 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.

[38] 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.

[39] E.L. Lehmann. *Testing Statistical Hypotheses*. Probability and Statistics Series. Wiley, 1986. ISBN: 9780471840831. URL: https://books.google.dk/books?id=jexQAAAAMAAJ.

[40] C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. 2nd. See Chapter 3 for the CramÃĺr-Rao inequality and its applications. New York: John Wiley & Sons, 1973. ISBN: 978-0-471-34969-5.

[41] J. Petersen. "The Missing MAss Problem on Galactic Scales." PhD thesis.

[42] 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.

[43] 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.

[44] M. Betancourt and Mark Girolami. "Hamiltonian Monte Carlo for Hierarchical Models." In: (Dec. 2013). DOI: 10.1201/b18502-5.

[45] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002.

Index