INTRODUCTION TO STATISTICS

THEORY, METHODS, AND APPLICATIONS

JONAS PETERSEN

This page is intentionally left blank

Contents

CHAPTER 1

Introduction

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, 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

This book has been shaped by many sources of inspiration. A few exercises are adapted from [1], the idea of presenting decision theory as a contest between "Robot vs. Nature" is inspired by [2], and the overall style has been influenced by works such as [1, 3, 4].

CHAPTER 2

Introduction to Set Theory

Set theory is a foundational branch of mathematics that provides the language and structure underlying much of modern mathematics, including probability theory. At its core, it studies sets—collections of distinct objects or elements—and the relationships between them. This chapter reviews the essential properties and operations of sets, laying the groundwork for the axiomatic development of probability theory and, ultimately, statistics.

Definition 1 (Membership)**.** *Given an object o and a set A, the notation $o \in A$ denotes that o is an element (or member) of A. If $o \notin A$, then o is not an element of A.*

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

1. *By listing their elements separated by commas, e.g.*

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

2. *By specifying a defining property of their elements, e.g.*

$$A = \{x \mid x \text{ is a natural number}\}. \tag{2}$$

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

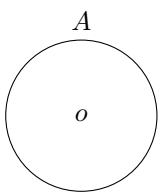

Figure 1: The graphical representation of a generic set A with generic element o.

Definition 3 (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 2.1.
Let $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, b_3\}$ then

$$A \times B = \{(a_1, b_1), (a_1, b_2), (a_1, b_3), (a_2, b_1), (a_2, b_2), (a_2, b_3)\}. \tag{3}$$

Definition 4 (Subset)**.** *The set A is called a subset of the 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.*

Definition 5 (Proper Subset)**.** *The set A is called a proper subset of the 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.*

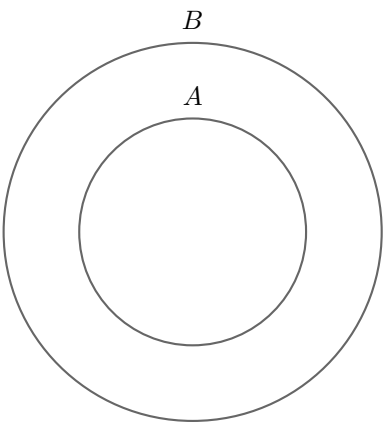

Figure 2: The graphical representation of $A \subset B$.

Example 2.2.
🍌, 🍎, and 🍆 are members (elements) of the set {🍌, 🍎, 🍆}, but are not subsets of it; in turn, the subsets, such as {🍌}, are not members of the set {🍌, 🍎, 🍆}.

Example 2.3.
Suppose $A = \{$🍌, 🍎, 🍆$\}$, then {🍌, 🍎} and {🍎} are proper subsets of A, meaning {🍌, 🍎}, {🍎} $\subset A$. {🍌, 🥕}, on the other hand, is not a subset of A, meaning {🍌, 🥕} $\not\subset A$.

Definition 6 (Empty Set)**.** *The empty set, denoted by* $\emptyset$ *or* $\{\}$*, is the set that contains no elements.*

Definition 7 (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 2.4.
The power set of the set $A = \{a_1, a_2, a_3\}$ *can be written*

$$2^A = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}\}. \tag{4}$$

Definition 8 (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 9 (Closure)**.** *The 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 10 (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 provides a graphical representation of* $A \cup B$*.*

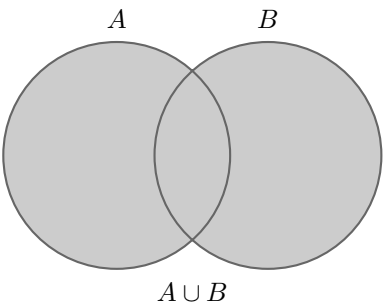

Figure 3: The graphical representation of $A \cup B$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 11 (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 12 (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 provides a graphical representation of $A \cap B$.*

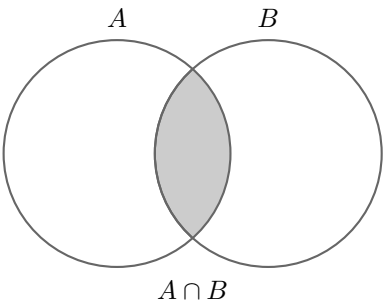

Figure 4: The graphical representation of $A \cap B$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 13 (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 14 (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 provides a graphical representation of $A \cap B = \emptyset$.*

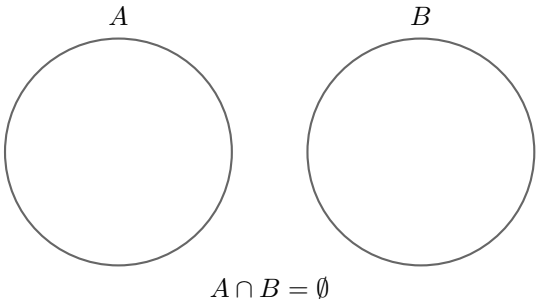

Figure 5: The graphical representation of $A \cap B = \emptyset$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 15 (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 provides a graphical representation of $(A \cap B)^c$.*

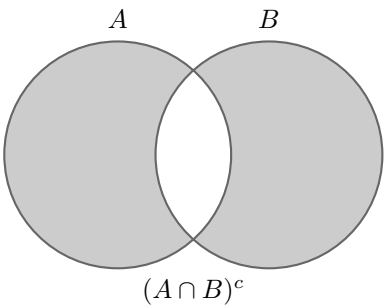

Figure 6: The graphical representation of $(A \cap B)^c$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 16 (Difference)**.** *The difference between sets 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 provides a graphical representation of $A \setminus B$ and $B \setminus A$.*

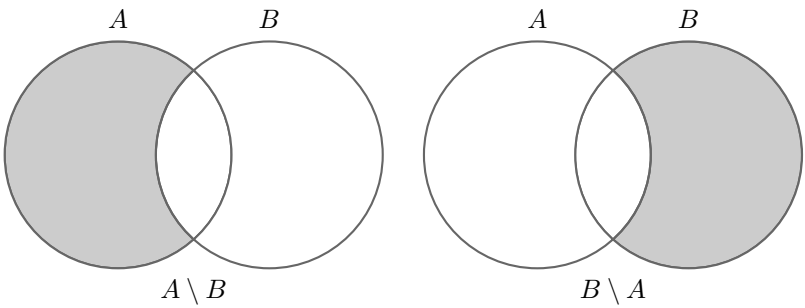

Figure 7: The graphical representation of $A \setminus B$ (left) and $B \setminus A$ (right). Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 17 (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 shows the symmetric difference between sets A and B.*

Definition 18 (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_i 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.

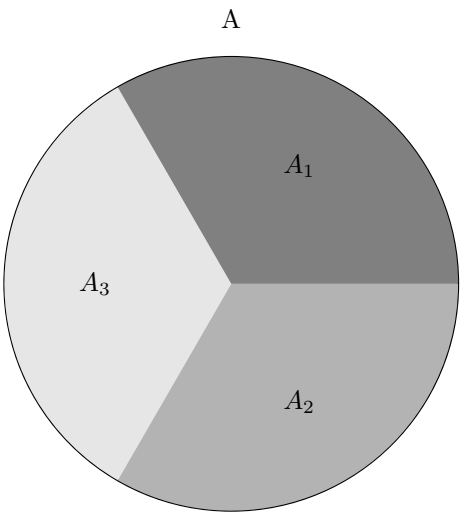

Figure 8: The graphical representation of $A = \{A_1, A_2, A_3\}$, where A_j are partitions.

CHAPTER 3

Introduction to Probability Theory

Probability theory is a foundational branch of mathematics that provides the formal framework for reasoning about uncertainty. At its core, it studies random experiments and the likelihood of their outcomes. This chapter reviews the essential principles and axioms of probability, laying the groundwork for statistical inference and decision-making under uncertainty.

Definition 19 (Sample Space)**.** *The sample space, denoted by Ω, is 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 (σ-algebra)**.** *A σ-algebra over a sample space Ω is a collection of subsets $\mathcal{G}$ of Ω that contains both $\emptyset$ and Ω, is closed under complementation (that is, if $E \in \mathcal{G}$ then $E^c \in \mathcal{G}$), and is closed under countable unions (and therefore also under countable intersections).*

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 trivial σ-algebra on Ω is given by

$$\mathcal{G}_{trivial} = \{\emptyset, \Omega\}. \tag{5}$$

In this case, the only events that can be described are the impossible event $\emptyset$ and the certain event Ω. For instance, the event of rolling an even number $E = \{\boxdot, \boxdot, \boxdot\}$ is not in $\mathcal{G}_{trivial}$.

Definition 22 (Borel σ-algebra)**.** *The Borel σ-algebra, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra on $\mathbb{R}$ that contains all open subsets of $\mathbb{R}$. Equivalently, $\mathcal{B}(\mathbb{R})$ is generated by the collection of open intervals $(a, b) \subset \mathbb{R}$. Thus, $\mathcal{B}(\mathbb{R})$ contains all sets that can be formed from open intervals through countable unions, intersections, and complements.*

Definition 23 (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 .*

Remark 1. *For a discrete sample space Ω, $\mathcal{F}$ is typically the power set of Ω. For a continuous sample space, $\mathcal{F}$ is typically the Borel σ-algebra, generated by open sets in $\mathbb{R}$ ($\mathcal{B}(\mathbb{R})$).*

Example 3.3.

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 is given by

$$\begin{aligned}\mathcal{F} &= \{\emptyset, \{\boxdot\}, \{\boxdot, \boxdot\}, \{\boxdot\}, \{\boxdot, \boxdot, \boxdot, \boxdot\}, \{\boxdot\}, \ldots, \Omega\} \\ &= 2^{\Omega}.\end{aligned} \tag{6}$$

Definition 24 (Measurable Space)**.** *A measurable space is a pair $(\Omega, \mathcal{F})$, where Ω is the sample space of a random experiment and $\mathcal{F}$ is the event space.*

Definition 25 (Measure)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, where Ω is the sample space and $\mathcal{F}$ is the event space. A measure μ is a set function*

$$\mu : \mathcal{F} \to [0, \infty] \tag{7}$$

that satisfies Axiom 1 (non-negativity) and Axiom 2 (additivity).

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

$$\mu(E) \geq 0 \quad \forall E \in \mathcal{F}. \tag{8}$$

Axiom 2 (Additivity)**.** *For any countable sequence of mutually exclusive events $E_1, E_2, \ldots \in \mathcal{F}$, the measure of their union is the sum of their individual measures, such that*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i) \quad \forall E_i \in \mathcal{F} \text{ where } \bigcap_{i=1}^{\infty} E_i = \emptyset. \tag{9}$$

Definition 26 (σ-finite Measure)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, where Ω is the sample space and $\mathcal{F}$ is the event space. A measure μ on $(\Omega, \mathcal{F})$ is called σ-finite if there exists a countable collection of sets $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ such that*

$$\Omega = \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \mu(A_i) < \infty \ \ \forall i \in \mathbb{N}. \tag{10}$$

Definition 27 (Measurable Function)**.** *Let $(\Omega, \mathcal{F})$ and $(\Omega_X, \mathcal{F}_X)$ be measurable spaces. A function*

$$X : \Omega \to \Omega_X \tag{11}$$

is said to be measurable if

$$X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{F}_X, \tag{12}$$

where

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}. \tag{13}$$

In words, the preimage of every measurable set in $\mathcal{F}_X$ is a measurable set in $\mathcal{F}$.

Definition 28 (Lebesgue Measure)**.** *The Lebesgue measure λ is a measure, in the sense of Definition 25, defined on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ such that for every interval $(a, b] \subseteq \mathbb{R}$,*

$$\lambda((a, b]) = b - a. \tag{14}$$

In higher dimensions, λ generalizes to $\mathbb{R}^n$, where it coincides with the usual notions of length, area, and volume.

Definition 29 (Counting Measure)**.** *Let $(\Omega, \mathcal{F})$ be a discrete measurable space, where Ω is the sample space and $\mathcal{F}$ is the event space. The counting measure ν is a measure, in the sense of Definition 25, defined on $\mathcal{F}$ such that for every event $E \in \mathcal{F}$,*

$$\nu(E) = |E|, \tag{15}$$

where $|E|$ denotes the cardinality of E (finite or countably infinite). In particular, for finite sets E, $\nu(E)$ equals the number of elements in E, and for countably infinite sets, $\nu(E) = \infty$.

Definition 30 (Probability Measure)**.** *Loosely speaking, probability can be regarded [4] as a measure of the size of an event relative to the sample space. Formally, a probability measure* $\mathbb{P}$ *is a measure, in accordance with Definition 25 (measure), defined on a measurable space* $(\Omega, \mathcal{F})$ *that, in addition to satisfying Axiom 1 (non-negativity) and Axiom 2 (additivity), obeys the normalization property*

$$\mathbb{P}(\Omega) = 1. \tag{16}$$

Definition 31 (Probability Space)**.** *A probability space is a triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where* Ω *is the sample space,* $\mathcal{F}$ *is the event space and* $\mathbb{P}$ *is a probability measure on the measurable space* $(\Omega, \mathcal{F})$*.*

Remark 2. *In continuous sample spaces, individual outcomes (single real numbers) have probability zero. Therefore, events are nontrivial subsets of the sample space, typically intervals or more general Borel sets. For example, the event*

$$E = (0.2, 0.5) \tag{17}$$

represents the outcome that the randomly chosen number lies between 0.2 *and* 0.5*.*

Remark 3. *The restriction to Borel sets in case of a continuous sample space (rather than all subsets of* $\mathbb{R}$*) is not arbitrary: it avoids paradoxical constructions such as non-measurable sets (e.g. the Vitali set), which cannot be consistently assigned a probability. This ensures that the probability measure is well defined for all events in* $\mathcal{F}$*.*

Example 3.4.
Consider choosing a real number uniformly at random from the interval $[0, 1]$*. Here the sample space is* $\Omega = [0, 1]$*. Unlike the discrete case, the event space* $\mathcal{F}$ *cannot simply be the power set of* $[0, 1]$*, since not all subsets admit a well-defined probability measure. Instead, the event space is chosen as the Borel* σ*-algebra* $\mathcal{B}([0, 1])$*, which includes sets such as open intervals* $(0.2, 0.5)$*, closed intervals* $[0, 0.1]$*, and countable unions and intersections thereof.*

Definition 32 (Independence)**.** *Events* E_1 *and* E_2 *in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *are said to be independent iff*

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2). \tag{18}$$

Definition 33 (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 as follows*

$$\mathbb{P}(E_1|E_2) \equiv \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}. \tag{19}$$

Definition 34 (Conditional Independence)**.** *Events E_1 and E_2 are conditionally independent given E_3 if*

$$\mathbb{P}(E_1 \cap E_2|E_3) = \mathbb{P}(E_1|E_3)\mathbb{P}(E_2|E_3). \tag{20}$$

Theorem 1 (Chain Rule)**.** *Let $E_1, E_2, \ldots, E_n \in \mathcal{F}$ be events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the chain rule states that*

$$\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) = \mathbb{P}(E_1) \prod_{j=2}^{n} \mathbb{P}(E_j \mid E_1 \cap \cdots \cap E_{j-1}). \tag{21}$$

Proof. From the definition of conditional probability in Definition 33

$$\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n) = \mathbb{P}(E_1|E_2 \cap \cdots \cap E_n)\mathbb{P}(E_2 \cap \cdots \cap E_n). \tag{22}$$

Using the definition of conditional probability again

$$\mathbb{P}(E_2 \cap \ldots \cap E_n) = \mathbb{P}(E_2| \ldots \cap E_n)\mathbb{P}(\cdots \cap E_n). \tag{23}$$

Continuing in this way, Theorem 1 follows. □

Remark 4. *Theorem 1 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 $E_1, E_2 \in \mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Bayes theorem can be formulated as follows*

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}. \tag{24}$$

Proof. Bayes theorem follows directly from applying Theorem 1 and applying the concept of symmetry as follows

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &= \mathbb{P}(E_1|E_2)\mathbb{P}(E_2) \\ &= \mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\end{aligned} \tag{25}$$

from which

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)} \tag{26}$$

which is Theorem 2. □

Theorem 3 (Law of Total Probability)**.** *Let* $\{E_1, E_2, \ldots, E_n\}$ *be a finite partition, in the sense of Definition 18, of the sample space* Ω *in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. Then, for any event* $A \in \mathcal{F}$*,*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap E_i). \tag{27}$$

Proof. Consider an event $A \in \mathcal{F}$ and a partition $\{E_1, E_2, \ldots E_n\}$ 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

$$\sum_i \mathbb{P}(A \cap E_i) = \mathbb{P}\Big(\bigcup_i (A \cap E_i)\Big). \tag{28}$$

$\bigcup_i (A \cap E_i)$ 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

$$\begin{aligned}\bigcup_i (A \cap E_i) &= (A, \bigcup_i E_i) \\ &= (A \cap \Omega) \\ &= A.\end{aligned} \tag{29}$$

Combining Equation 28 and Equation 29 then yields

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap E_i) \tag{30}$$

which is Theorem 3. □

Example 3.5.

For the roll with the fair die considered in Example 3.1, Example 3.2 and Example 3.3, the sample space is $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. *Let* $E_1 = \{\boxdot, \boxdot, \boxdot\}$ *and* $E_2 = \{\boxdot\}$ *be two events, then from Definition 33*

$$\begin{aligned} \mathbb{P}(E_1|E_2) &= \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \\ &= 1 \end{aligned} \tag{31}$$

where $\mathbb{P}(E_1 \cap E_2) = \frac{1}{6}$ *since* $E_1 \cap E_2 = E_2 = \{\boxdot\}$ *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 35 (Random Variable)**.** *A random variable* X *is a measurable function in the sense of Definition 27,*

$$X : \Omega \to \Omega_X \tag{32}$$

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.

Remark 5. *Random variables provide a numerical representation of the outcomes of a random experiment. They are classified as either discrete, when* Ω_X *is countable, or continuous, when* Ω_X *is uncountable, often modeled as an interval of* $\mathbb{R}$.

Definition 36 (Image Measure)**.** *Let*

$$X : \Omega \to \Omega_X \tag{33}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$. *Then [5]*

$$\mathbb{P} \circ X^{-1} : \mathcal{F}_X \to [0, 1] \tag{34}$$

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* $\mathbb{P}$.

Remark 6. *Theorem 3 extends naturally from a finite or countable partition of the sample space to the case where the partition is induced by a random variable* X. *Let*

$$X : \Omega \to \Omega_X \tag{35}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$*. Then, for any event* $A \in \mathcal{F}$*, Theorem 3 can be written rigorously in terms of the image measure* $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ *as*

$$\mathbb{P}(A) = \int_{\Omega_X} \mathbb{P}(A \mid X^{-1}(\{x\}))\, d\mathbb{P}_X(x), \tag{36}$$

where $\mathbb{P}(A \mid X^{-1}(\{x\}))$ *is the conditional probability of* A *given the event* $X^{-1}(\{x\}) \subseteq \Omega$*.*

Definition 37 (Expected value)**.** *Let*

$$X : \Omega \to \Omega_X \tag{37}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$*, and let*

$$\mathbb{P}_X = \mathbb{P} \circ X^{-1} \tag{38}$$

be the image measure of X *on* $(\Omega_X, \mathcal{F}_X)$*. The expected value of* X*, denoted by* $\mathbb{E}_X[X]$*, is defined as follows*

$$\mathbb{E}_X[X] \equiv \int_{\Omega_X} x d\mathbb{P}_X(x). \tag{39}$$

Theorem 4 (Non-negativity of expected value)**.** *Let*

$$X : \Omega \to \Omega_X \tag{40}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$*.* $X \geq 0 \Rightarrow \mathbb{E}_X[X] \geq 0$*.*

Proof. From Definition 37

$$\mathbb{E}_X[X] = \int_{\Omega_X} x d\mathbb{P}_X(x), \tag{41}$$

and if $x \geq 0$ for all $x \in \Omega_X$, the integral of a non-negative function with respect to a measure is non-negative. Hence, $\mathbb{E}_X[X] \geq 0$. □

Theorem 5 (Linearity of expected value)**.** *Let*

$$X : \Omega \to \Omega_X \tag{42}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$*. The expected value is a linear operator meaning* $\mathbb{E}_X[a + X] = a + \mathbb{E}_X[X]$ *and* $\mathbb{E}_X[aX] = a\mathbb{E}_X[X]$ *for any constant* a*.*

Proof. From Definition 37

$$\begin{aligned}\mathbb{E}_X[a+X] &= \int_{\Omega_X}(a+x)d\mathbb{P}_X(x)\\ &= a\int_{\Omega_X}d\mathbb{P}_X(x)+\int_{\Omega_X}xd\mathbb{P}_X(x)\\ &= a+\mathbb{E}_X[X],\end{aligned} \tag{43}$$

since $\mathbb{P}_X(\Omega_X)=1$. Similarly,

$$\begin{aligned}\mathbb{E}_X[aX] &= \int_{\Omega_X}axd\mathbb{P}_X(x)\\ &= a\int_{\Omega_X}xd\mathbb{P}_X(x)\\ &= a\mathbb{E}_X[X].\end{aligned} \tag{44}$$

□

Remark 7 (Law of the Unconscious Statistician)**.** *Let*

$$X:\Omega\to\Omega_X \tag{45}$$

be a random variable from the probability space $(\Omega,\mathcal{F},\mathbb{P})$ *to a measurable space* $(\Omega_X,\mathcal{F}_X)$*, and let*

$$g:\Omega_X\to\mathbb{R} \tag{46}$$

be a generic measurable function. Denote the image measure of X *by* $\mathbb{P}_X$*. Then*

$$\mathbb{E}_X[g(X)]\equiv\int_{\Omega_X}g(x)d\mathbb{P}_X(x). \tag{47}$$

Definition 38 (Variance)**.** *Let*

$$X:\Omega\to\Omega_X \tag{48}$$

be a real-valued random variable from the probability space $(\Omega,\mathcal{F},\mathbb{P})$ *to a measurable space* $(\Omega_X,\mathcal{F}_X)$*. The variance of* X*, denoted by* $\mathrm{Var}_X[X]$*, is defined as follows*

$$\begin{aligned}\mathrm{Var}_X[X] &\equiv \mathbb{E}_X[(X-\mathbb{E}_X[X])^2]\\ &= \mathbb{E}_X[X^2]-\mathbb{E}_X[X]^2.\end{aligned} \tag{49}$$

Theorem 6 (Markov's Inequality)**.** *Let*

$$X : \Omega \to \Omega_X$$

be a non-negative random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$*, and let* $a > 0$*. Then*

$$\mathbb{P}_X([a, \infty)) \leq \frac{\mathbb{E}_X[X]}{a}. \tag{50}$$

Proof. Let $1_{[a,\infty)}$ denote the indicator of the event $\{x \in \Omega_X | x \geq a\}$. Since $X(\omega) \geq 0$ and $a > 0$,

$$a\, 1_{[a,\infty)}(X(\omega)) \leq X(\omega), \quad \forall \omega \in \Omega. \tag{51}$$

Taking expectations with respect to $\mathbb{P}_X$ and using linearity,

$$a\, \mathbb{E}_X[1_{[a,\infty)}] \leq \mathbb{E}_X[X]. \tag{52}$$

By definition of the image measure,

$$\begin{aligned} \mathbb{E}_X[1_{[a,\infty)}] &= \int 1_{[a,\infty)}(x) d\mathbb{P}_X(x) \\ &= \mathbb{P}_X([a, \infty)). \end{aligned} \tag{53}$$

Hence,

$$a\mathbb{P}_X([a, \infty)) \leq \mathbb{E}_X[X], \tag{54}$$

and dividing both sides by a yields the inequality. □

Definition 39 (Probability Density with Respect to a Measure)**.** *Let*

$$X : \Omega \to \Omega_X \tag{55}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to the measurable space* $(\Omega_X, \mathcal{F}_X)$*. Let* μ *be a* σ*-finite measure on* $(\Omega_X, \mathcal{F}_X)$*, in the sense of Definition 26. If the image measure* $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ *is absolutely continuous with respect to* μ*, then by the Radon-Nikodym theorem [6] there exists a measurable function*

$$f_X = \frac{d\mathbb{P}_X}{d\mu}, \tag{56}$$

called the probability density of X *with respect to* μ*, such that*

$$\mathbb{P}_X(B) = \int_B f_X(x) d\mu(x), \quad \forall B \in \mathcal{F}_X. \tag{57}$$

Moreover, since $\mathbb{P}_X$ is a probability measure, the density satisfies

$$f_X(x) \geq 0, \quad \textit{and} \quad \int_{\Omega_X} f_X(x) d\mu(x) = 1. \tag{58}$$

Remark 8 (Probability Density Function)**.** *If $\mu = \lambda$ is the Lebesgue measure (Definition 28) on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the probability density f_X is called the probability density function (PDF). For $\mu = \lambda$, Definition 39 gives*

$$\mathbb{P}_X(B) = \int_B f_X(x) d\lambda(x), \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{59}$$

Remark 9 (Probability Mass Function)**.** *If $\mu = \nu$ is the counting measure (Definition 29) on Ω_X, then the probability density f_X is called the probability mass function (PMF). For $\mu = \nu$, Definition 39 gives*

$$\begin{aligned} \mathbb{P}_X(B) &= \int_B f_X(x) d\nu(x) \\ &= \sum_{x \in B} f_X(x). \end{aligned} \tag{60}$$

In particular, for a singleton $B = \{x\}$,

$$\mathbb{P}_X(\{x\}) = f_X(x). \tag{61}$$

Remark 10 (Expected value of a discrete random variable)**.** *If X is a discrete random variable with probability mass function f_X, then the expected value reduces to*

$$\mathbb{E}_X[X] = \sum_{x \in \Omega_X} x f_X(x). \tag{62}$$

Equation 62 follows directly from Definition 37, since the image measure $\mathbb{P}_X$ is concentrated on singletons $\{x\}$ in the discrete case.

Remark 11 (Expected value of a continuous random variable)**.** *Let X be a continuous random variable with PDF f_X on $\Omega_X \subseteq \mathbb{R}$. From Definition 37 and Definition 39*

$$\mathbb{E}_X[X] = \int_{\Omega_X} x f_X(x) d\lambda(x), \tag{63}$$

where λ denotes the Lebesgue measure on $\mathbb{R}$. In practice, it is customary to write $d\lambda(x)$ simply as dx, so that

$$\mathbb{E}_X[X] = \int_{\Omega_X} x f_X(x) dx. \tag{64}$$

Here, dx is understood as integration with respect to the Lebesgue measure.

Example 3.6.

Let X be a continuous random variable with PDF f_X on $\Omega_X \subseteq \mathbb{R}$. For the interval (event) $[a,b] \subseteq \Omega_X$,

$$\begin{aligned}\mathbb{P}(X^{-1}([a,b])) &= \mathbb{P}_X([a,b]) \\ &= \int_{[a,b]} f_X(x) d\lambda(x) \\ &= \int_a^b f_X(x) dx.\end{aligned} \tag{65}$$

Definition 40 (Joint Probability Measure)**.** *Let*

$$X : \Omega \to \Omega_X, \quad Y : \Omega \to \Omega_Y \tag{66}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$. The joint probability measure of X and Y is the image measure

$$\mathbb{P}_{X,Y} = \mathbb{P} \circ (X,Y)^{-1} \tag{67}$$

defined on the measurable space

$$(\Omega_{X_1} \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y). \tag{68}$$

All probability measures related to the random variables can be derived from the joint probability measure via Theorem 3.

Remark 12. *Let*

$$X : \Omega \to \Omega_X, \quad Y : \Omega \to \Omega_Y \tag{69}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$. Suppose $A \in \mathcal{F}_X$ and let $\mathbb{P}_{X,Y}$ denote the joint probability measure on the measurable space $(\Omega_{X_1} \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$, then from Theorem 3

$$\mathbb{P}_X(A) = \mathbb{P}_{X,Y}(A \times \Omega_Y). \tag{70}$$

Theorem 7 (Law of total expectation)**.** *Let*

$$X : \Omega \to \Omega_X, \quad Y : \Omega \to \Omega_Y \tag{71}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$. Let $\mathbb{P}_{X,Y}$ denote the joint probability measure on the measurable space $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$. Then

$$\mathbb{E}_X[X] = \mathbb{E}_Y\big[\mathbb{E}_{X|Y}[X \mid Y]\big]. \tag{72}$$

Proof. Let

$$X : \Omega \to \Omega_X, \quad Y : \Omega \to \Omega_Y \tag{73}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$. Let $\mathbb{P}_{X,Y}$ denote the joint probability measure on the measurable space $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$. By Fubini's theorem and Definition 33,

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} x d\mathbb{P}_X(x) \\
&= \int_{\Omega_X} \int_{\Omega_Y} x d\mathbb{P}_{X,Y}(x, y) \\
&= \int_{\Omega_Y} \int_{\Omega_X} x d\mathbb{P}_{X|Y}(x) d\mathbb{P}_Y(y) \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X \mid Y]].
\end{aligned} \tag{74}$$

or equivalently in terms of the probability densities

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} x f_X(x) d\mu_X(x) \\
&= \int_{\Omega_Y} \int_{\Omega_X} x f_{X,Y}(x, y) d\mu_X(x) d\mu_Y(y) \\
&= \int_{\Omega_Y} f_Y(y) \left(\int_{\Omega_X} x f_{X|Y}(x \mid y) d\mu_X(x) \right) d\mu_Y(y) \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X \mid Y]].
\end{aligned} \tag{75}$$

□

Theorem 8 (Expectation of product 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 the random variables are independent the expectation can be written*

$$\mathbb{E}_{X,Y}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]. \tag{76}$$

Proof. If the random variables are independent, then according to Definition 32

$$\mathbb{P}_{X,Y}(\{x, y\}) = \mathbb{P}_X(\{x\})\mathbb{P}_Y(\{y\}) \tag{77}$$

meaning

$$\begin{aligned}\mathbb{E}_{X,Y}[XY] &= \int_{\Omega_X}\int_{\Omega_Y} xyd\mathbb{P}_{X,Y}(x,y) \\ &= \int_{\Omega_X} xd\mathbb{P}_X(x)\int_{\Omega_Y} yd\mathbb{P}_Y(y) \\ &= \mathbb{E}_X[X]\mathbb{E}_Y[Y].\end{aligned} \tag{78}$$

□

Definition 41 (Covariance)**.** *Let $X : \Omega \to \Omega_X$ and $Y : \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the covariance of X and Y, denoted by $\mathrm{Cov}_{X,Y}[X,Y]$, is defined as follows*

$$\begin{aligned}\mathrm{Cov}_{X,Y}[X,Y] &= \mathbb{E}_{X,Y}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y],\end{aligned} \tag{79}$$

Theorem 9 (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*

$$\mathrm{Cov}_{X,Y}[X,Y] = 0. \tag{80}$$

Proof. Using Theorem 8 in Definition 41 yields $\mathrm{Cov}_{X,Y}[X,Y] = 0$. □

Definition 42 (Correlation)**.** *Let $X : \Omega \to \Omega_X$ and $Y : \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The correlation between X and Y, denoted by $\mathrm{Corr}[X,Y]$, is defined as*

$$\begin{aligned}\mathrm{Corr}_{X,Y}[X,Y] &= \frac{\mathrm{Cov}_{X,Y}[X,Y]}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_Y[Y]}} \\ &= \frac{\mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2)\,(\mathbb{E}_Y[Y^2] - \mathbb{E}_Y[Y]^2)}}.\end{aligned} \tag{81}$$

Remark 13. *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 $\mathrm{Corr}_{X,Y}[X,Y]$, 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 43 (Change of Variables for PDFs)**.** *Let $X : \Omega \to \Omega_X$ be a continuous random variable with probability density function (PDF) f_X, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose $Y = g(X)$, where g is a continuous and differentiable function with differentiable inverse g^{-1}. Then the PDF of Y, denoted f_Y, is given by the change of variables formula [3]*

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right|, \quad y \in \Omega_Y, \tag{82}$$

where $\Omega_Y = g(\Omega_X)$ is the codomain of Y.

Example 3.7.

Let $X : \Omega \to \Omega_X$ and $Y : \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Using Definition 38 and Definition 41, the variance of a sum can be written

$$\begin{aligned} \mathrm{Var}_{X,Y}[X+Y] &= \mathbb{E}_{X,Y}[(X+Y-\mathbb{E}_{X,Y}[X+Y])^2] \\ &= \mathbb{E}_X[(X-\mathbb{E}_X[X])^2] + \mathbb{E}_Y[(Y-\mathbb{E}_Y[Y])^2] \\ &\quad + 2\mathbb{E}_{X,Y}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= \mathrm{Var}_X[X] + \mathrm{Var}_Y[Y] + 2\,\mathrm{Cov}_{X,Y}[X,Y]. \end{aligned} \tag{83}$$

Example 3.8.

Let $X : \Omega \to \Omega_X$ be a continuous random variable with probability density function (PDF) f_X, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose

$$\begin{aligned} Y &= g(X) \\ &= aX + b, \end{aligned} \tag{84}$$

where $a \neq 0$ and b are constants. The inverse function is

$$g^{-1}(y) = \frac{y-b}{a}. \tag{85}$$

Using Definition 43, the PDF of Y is

$$\begin{aligned} f_Y(y) &= f_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= f_X\left(\frac{y-b}{a}\right) \left| \frac{1}{a} \right|. \end{aligned} \tag{86}$$

Hence,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right). \tag{87}$$

Example 3.9.

Let $X : \Omega \to \Omega_X$ be a continuous random variable with probability density function (PDF) f_X, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose $f_X(x) \propto \text{const}$, and let

$$\begin{aligned} Y &= g(X) \\ &= \frac{e^X}{1+e^X}. \end{aligned} \tag{88}$$

The inverse function is

$$g^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{89}$$

Using Definition 43, the PDF of Y is

$$\begin{aligned} f_Y(y) &= f_X\big(g^{-1}(y)\big)\left|\frac{d}{dy}g^{-1}(y)\right| \\ &= \text{const}\cdot\left|\frac{d}{dy}\ln\left(\frac{y}{1-y}\right)\right| \\ &= \text{const}\cdot\frac{1}{y(1-y)}, \quad y \in (0,1). \end{aligned} \tag{90}$$

Theorem 10 (Error Propagation)**.** *Let*

$$X_1 : \Omega \to \Omega_{X_1}, \ldots, X_n : \Omega \to \Omega_{X_n} \tag{91}$$

be continuous random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let

$$g : \Omega_{X_1} \times \cdots \times \Omega_{X_n} \to \mathbb{R} \tag{92}$$

be a differentiable function of these variables. Denote for shorthand

$$X = (X_1, \ldots, X_n) \tag{93}$$

and

$$\mathbb{E}_X[X] = (\mathbb{E}_{X_1}[X_1], \ldots, \mathbb{E}_{X_n}[X_n]). \tag{94}$$

Then the variance of $g(X)$, which quantifies the uncertainty in g due to the uncertainties in $X_1, \ldots, X_n$, satisfies the first-order approximation

$$\begin{aligned}\mathrm{Var}_X[g(X)] = &\sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \\ &+ \sum_{i \neq j} \frac{\partial g(X)}{\partial X_i} \frac{\partial g(X)}{\partial X_j}\bigg|_{X=\mathbb{E}_X[X]} \mathrm{Cov}_{X_i,X_j}[X_i, X_j] \\ &+ \mathcal{O}(\|X - \mathbb{E}_X[X]\|^3).\end{aligned} \tag{95}$$

Proof. $g(X)$ can be written as a Taylor expansion around $\mathbb{E}_X[X]$ as follows

$$\begin{aligned}g(X) =& g(\mathbb{E}_X[X]) + \sum_{i=1}^{n} \frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} (X_i - \mathbb{E}_{X_i}[X_i]) \\ &+ \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2).\end{aligned} \tag{96}$$

Consequently

$$\begin{aligned}\mathbb{E}_X[g(X)] &= g(\mathbb{E}_X[X]) + \sum_{i=1}^{n} \frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \underbrace{\mathbb{E}_{X_i}[X_i - \mathbb{E}_{X_i}[X_i]]}_{\to 0} \\ &\quad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2) \\ &= g(\mathbb{E}_X[X]) + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2)\end{aligned} \tag{97}$$

meaning the variance of g can be approximated as follows

$$\begin{aligned}\mathrm{Var}_X[g(X)] &= \mathbb{E}_X\left[(g(X) - \mathbb{E}_X[g(X)])^2\right] \\ &= \mathbb{E}_X\Bigg[\Bigg(\sum_{i=1}^{n} (X_i - \mathbb{E}_{X_i}[X_i]) \frac{\partial g}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \\ &\qquad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2)\Bigg)^2\Bigg] \\ &= \sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \\ &\quad + \sum_{i \neq j} \frac{\partial g(X)}{\partial X_i} \frac{\partial g(X)}{\partial X_j}\bigg|_{X=\mathbb{E}_X[X]} \mathrm{Cov}_{X_i,X_j}[X_i, X_j] \\ &\quad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^3).\end{aligned} \tag{98}$$

□

Remark 14 (Error propagation for independent Random Variables)**.** *Let*

$$X_1 : \Omega \to \Omega_{X_1}, \ldots, X_n : \Omega \to \Omega_{X_n} \tag{99}$$

be continuous random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$*. If the random variables* $X_1, \ldots, X_n$ *are independent, then according to Theorem 9* $\mathrm{Cov}_{X_i,X_j}[X_i, X_j] = 0 \ \forall i \neq j$*. In this case, Theorem 10 simplifies to*

$$\mathrm{Var}_X[g(X)] \approx \sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\Big|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \tag{100}$$

where

$$X = (X_1, \ldots, X_n) \tag{101}$$

and

$$\mathbb{E}_X[X] = (\mathbb{E}_{X_1}[X_1], \ldots, \mathbb{E}_{X_n}[X_n]). \tag{102}$$

Equation 100 is the commonly used form of the linear error-propagation formula for independent variables.

Example 3.10.

A company produces square plates with dimensions characterized by two independent random variables

$$X \sim \mathrm{Norm}(2m, (0.01m)^2), \quad Y \sim \mathrm{Norm}(3m, (0.02m)^2). \tag{103}$$

The variance of the area XY *can be determined exactly from Theorem 10*

$$\begin{aligned} \mathrm{Var}_{X,Y}[XY] &= \mathbb{E}_{X,Y}[(XY)^2] - (\mathbb{E}_{X,Y}[XY])^2 \\ &= (\mathrm{Var}_X[X] + \mathbb{E}_X[X]) \Big(\mathrm{Var}_Y[Y] + \mathbb{E}_Y[Y]) - \mathbb{E}_X[X]^2 \mathbb{E}_Y[Y]^2 \\ &= \mathbb{E}_Y[Y]^2 \, \mathrm{Var}_X[X] + \mathbb{E}_X[X]^2 \, \mathrm{Var}_Y[Y] + \mathrm{Var}_X[X] \, \mathrm{Var}_Y[Y] \end{aligned} \tag{104}$$

where Theorem 8 has been applied. Via the linear approximation from Remark 14 the variance of the area can be approximated as follows

$$\begin{aligned} \mathrm{Var}_{X,Y}[XY]|_{\textit{linear approx.}} &\approx \sum_{i=X,Y} \left(\frac{\partial (XY)}{\partial i}\Big|_{X=\mathbb{E}_X[X], Y=\mathbb{E}_Y[Y]} \right)^2 \mathrm{Var}_i[i] \\ &= \mathbb{E}_Y[Y]^2 \, \mathrm{Var}_X[X] + \mathbb{E}_X[X]^2 \, \mathrm{Var}_Y[Y] \end{aligned}$$

(105)

Comparing Equation 104 and Equation 105 the relative difference can be written

$$\begin{aligned}\frac{\mathrm{Var}_{X,Y}[XY]|_{\textit{linear approx.}} - \mathrm{Var}_{X,Y}[XY]}{\mathrm{Var}_{X,Y}[XY]} &= -\frac{\mathrm{Var}_X[X]\,\mathrm{Var}_Y[Y]}{\mathrm{Var}_{X,Y}[XY]} \\ &\simeq -1.6 \cdot 10^{-5}.\end{aligned} \tag{106}$$

Example 3.11.
Consider a probability space describing two children with unknown sexes. Let

$$\Omega_{child\ 1} = \{\text{👦}, \text{👧}\}, \quad \Omega_{child\ 2} = \{\text{👦}, \text{👧}\}, \tag{107}$$

and define

$$\Omega = \Omega_{child\ 1} \times \Omega_{child\ 2} = \{(\text{👦},\text{👦}), (\text{👦},\text{👧}), (\text{👧},\text{👦}), (\text{👧},\text{👧})\}. \tag{108}$$

Define the random variables

$$B : \Omega \to \{0,1,2\}, \quad G : \Omega \to \{0,1,2\}, \tag{109}$$

where $B(\omega)$ and $G(\omega)$ denote the number of boys and girls in outcome $\omega \in \Omega$. The joint probability mass function of (B, G) is, by Definition 39,

$$f_{B,G}(b,g) = \mathbb{P}_{B,G}(\{(b,g)\}) = \mathbb{P}(\{\omega \in \Omega | B(\omega) = b,\, G(\omega) = g\}). \tag{110}$$

For instance,

$$f_{B,G}(1,1) = \mathbb{P}(\{(\text{👦},\text{👧}), (\text{👧},\text{👦})\}) = \frac{1}{2}. \tag{111}$$

Let $A \subseteq \Omega_B \times \Omega_G$ denote the event "at least one boy":

$$A = \{(b,g) \in \Omega_B \times \Omega_G \mid b \geq 1\}. \tag{112}$$

Using Definition 33 the conditional probability of exactly one girl given at least one boy is (see Remark 15)

$$\mathbb{P}_{B,G}(\{(1,1)\} \mid A) = \frac{\mathbb{P}_{B,G}(\{(1,1)\} \cap A)}{\mathbb{P}_{B,G}(A)}. \tag{113}$$

From the PMF,

$$\mathbb{P}_{B,G}(\{(1,1)\} \cap A) = f_{B,G}(1,1) = \frac{1}{2}, \tag{114}$$

and from Theorem 3

$$\mathbb{P}_{B,G}(A) = \sum_{g} \sum_{b \geq 1} f_{B,G}(b,g) = f_{B,G}(1,1) + f_{B,G}(2,0) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}. \tag{115}$$

Hence,

$$\mathbb{P}_{B,G}(\{(1,1)\} \mid A) = \frac{2}{3}. \tag{116}$$

Remark 15 (PMF vs. image measure on events)**.** *Let B and G be discrete random variables with joint PMF $f_{B,G}$ and image measure $\mathbb{P}_{B,G}$. From Definition 39, the PMF is defined only for single points $(b, g) \in \Omega_B \times \Omega_G$:*

$$f_{B,G}(b, g) = \mathbb{P}_{B,G}(\{(b, g)\}). \tag{117}$$

The image measure $\mathbb{P}_{B,G}$, however, is defined on all measurable subsets in the event space $A \in \mathcal{F}_B \otimes \mathcal{F}_G$. Hence, expressions of the form

$$\mathbb{P}_{B,G}(\{(b, g)\} \cap A) \tag{118}$$

are valid, since $\{(b, g)\} \cap A$ is an element of $\mathcal{F}_B \otimes \mathcal{F}_G$. In contrast, writing

$$f_{B,G}(b, g \cap A) \tag{119}$$

(or similarly) is not valid, because $f_{B,G}$ is defined only on individual points (b, g), not on sets or intersections. The PMF cannot take an event as its argument; only the image measure $\mathbb{P}_{B,G}$ can.

Example 3.12.

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

$$\mathbb{P}(E) = 0.01. \tag{120}$$

The prosecutor claims: "There is a 1% chance that the defendant would have the blood type found at the crime scene 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

$$\mathbb{P}(E \mid \text{innocent}) = 0.01 \Rightarrow \mathbb{P}(\text{guilty} \mid E) = 0.99. \tag{121}$$

To investigate this claim, use Theorem 2 to write

$$\begin{aligned} \mathbb{P}(E \mid \text{innocent}) &= \frac{\mathbb{P}(E \cap \text{innocent})}{\mathbb{P}(\text{innocent})} \\ &= \frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})} \mathbb{P}(E). \end{aligned} \tag{122}$$

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,*

$$\frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})} = \frac{\frac{M-1}{M}}{\frac{N-1}{N}}, \tag{123}$$

which approaches 1 *in the limit* $N, M \to \infty$. *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 33, as follows*

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(\text{guilty} \cap E) + \mathbb{P}(\text{innocent} \cap E)}{\mathbb{P}(E)}. \tag{124}$$

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

$$\mathbb{P}(\text{guilty} \cap E) + \mathbb{P}(\text{innocent} \cap E) = \mathbb{P}(E), \tag{125}$$

and thereby

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = 1. \tag{126}$$

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,*

$$\mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(E \mid \text{innocent})\,\mathbb{P}(\text{innocent})}{\mathbb{P}(E)} \tag{127}$$

From equation (127), *it is clear that in general*

$$\mathbb{P}(E \mid \text{innocent}) \neq \mathbb{P}(\text{innocent} \mid E), \tag{128}$$

and so even if $\mathbb{P}(E \mid \text{innocent}) = 0.01$, *the prosecutor's claim (Equation 121) is not true.*

Example 3.13.

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. Using Theorem 2 and Theorem 3

$$\begin{aligned}\mathbb{P}(s|p) &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p)} \\ &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p|s)\mathbb{P}(s)+\mathbb{P}(p|h)\mathbb{P}(h)}\end{aligned} \tag{129}$$

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*

$$\mathbb{P}(s|p) \simeq 0.0098. \tag{130}$$

Example 3.14.

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*

$$z^* = \underset{z}{\operatorname{argmax}}(\mathbb{P}(z|o_3 \cap c_1)). \tag{131}$$

Since the host cannot open the door containing the prize,

$$\mathbb{P}(z_3|o_3 \cap c_1) = 0 \tag{132}$$

and only $\mathbb{P}(z_1|o_3 \cap c_1)$ *and* $\mathbb{P}(z_2|o_3 \cap c_1)$ *will have to be considered. Using Theorem 2*

$$\mathbb{P}(z_1|o_3 \cap c_1) = \frac{\mathbb{P}(o_3|c_1 \cap z_1)\mathbb{P}(c_1 \cap z_1)}{\mathbb{P}(o_3 \cap c_1)} \tag{133}$$

where from Theorem 3

$$\begin{aligned}\mathbb{P}(o_3 \cap c_1) &= \sum_i \mathbb{P}(o_3 \cap c_1 \cap z_i) \\ &= \mathbb{P}(o_3 \cap c_1 \cap z_1) + \mathbb{P}(o_3 \cap c_1 \cap z_2) + \mathbb{P}(o_3 \cap c_1 \cap z_3) \\ &= \mathbb{P}(o_3|c_1 \cap z_1)\mathbb{P}(c_1 \cap z_1) + \mathbb{P}(o_3|c_1 \cap z_2)\mathbb{P}(c_1 \cap z_2) \\ &\quad + \mathbb{P}(o_3|c_1 \cap z_3)\mathbb{P}(c_1 \cap z_3).\end{aligned} \tag{134}$$

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

$$\begin{aligned}\mathbb{P}(z_1|o_3 \cap c_1) &= \frac{\mathbb{P}(o_3|c_1 \cap z_1)}{\sum_i \mathbb{P}(o_3|c_1 \cap z_i)} \\ &= \frac{1}{3}.\end{aligned} \tag{135}$$

Similarly

$$\begin{aligned}\mathbb{P}(z_2|o_3 \cap c_1) &= \frac{\mathbb{P}(o_3|c_1 \cap z_2)}{\sum_i \mathbb{P}(o_3|c_1 \cap z_i)} \\ &= \frac{2}{3}.\end{aligned} \tag{136}$$

Since $\mathbb{P}(z_2|o_3 \cap c_1) > \mathbb{P}(z_1|o_3 \cap c_1) > \mathbb{P}(z_3|o_3 \cap c_1)$, *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* $\mathbb{P}(z_1|o_3 \cap c_1) = \frac{1}{3}$. *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 3.15.

Let $X : \Omega \to \Omega_X$ *and* $Y : \Omega \to \Omega_Y$ *be continuous random variables defined on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and suppose* $X \sim \text{Unif}(a = -1, b = 1)$ *and* $Y = X^2$. *Clearly* Y *is dependent on* X *(in fact* Y *is uniquely determined by* X*). Show that* $\text{Corr}_{X,Y}[X, Y] = 0$.

Since $Y = X^2$ *is uniquely determined by* X,

$$\text{Corr}_{X,Y}[X, Y] = \text{Corr}_X[X, X^2]. \tag{137}$$

Using Definition 42 and Definition 41

$$\begin{aligned}\mathrm{Corr}_X[X, X^2] &= \frac{\mathrm{Cov}_X[X, X^2]}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_X[X^2]}} \\ &= \frac{\mathbb{E}_X[X^3] - \mathbb{E}_X[X]\mathbb{E}_X[X^2]}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_X[X^2]}}\end{aligned} \tag{138}$$

In this case for the nominator

$$\begin{aligned}\mathrm{Cov}_X[X, X^2] &= \int_{\Omega_X} x^3 f_X(x)dx - \int_{\Omega_X} x' f_X(x')dx' \int_{\Omega_X} x''^2 f_X(x'')dx'' \\ &= \frac{1}{b-a}\int_a^b x^3 dx - \frac{1}{(b-a)^2}\int_a^b x'dx' \int_a^b x''^2 dx'' \\ &= \frac{1}{12}(a-b)^2(a+b) \\ &= 0\end{aligned} \tag{139}$$

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

$$\begin{aligned}\mathrm{Var}_X[X]\,\mathrm{Var}_X[X^2] &= \big(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2\big)\big(\mathbb{E}_X[X^4] - \mathbb{E}_X[X^2]^2\big) \\ &= \frac{1}{540}(b-a)^4(4a^2 + 7ab + 4b^2) \\ &\neq 0.\end{aligned} \tag{140}$$

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

Example 3.16.

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

1. *Show $Y \sim \mathrm{Norm}(\mu = 0, \sigma^2 = 1)$.*

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

$$\begin{aligned}\mathbb{E}_Y[Y] &= \mathbb{E}_{W,X}[WX] \\ &= \mathbb{E}_W[W]\cancelto{0}{\mathbb{E}_X[X]} \\ &= 0.\end{aligned} \tag{141}$$

The variance

$$\begin{aligned}\mathrm{Var}_Y[Y] &= \mathbb{E}_Y[Y^2] - \cancelto{0}{\mathbb{E}_Y[Y]^2} \\ &= \mathbb{E}_{W,X}[W^2X^2] \\ &= \mathbb{E}_W[W^2]\mathbb{E}_X[X^2] \\ &= \mathbb{E}_W[W^2]\,\mathrm{Var}_X[X]\end{aligned} \tag{142}$$

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

$$\begin{aligned}\mathbb{E}_W[W^2] &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 f_W(w_i) \\ &= \frac{1}{2}[(-1)^2\frac{1}{2} + 1^2\frac{1}{2}] \\ &= 1\end{aligned} \tag{143}$$

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

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

$$\begin{aligned}\mathrm{Cov}_{X,Y}[X,Y] &= \mathrm{Cov}_{X,W}[X, WX] \\ &= \mathbb{E}_{X,W}[WX^2] - \mathbb{E}_X[X]\mathbb{E}_{X,W}[WX] \\ &= \mathbb{E}_W[W]\mathbb{E}_X[X^2] - \mathbb{E}_W[W]\mathbb{E}_X[X]^2 \\ &= \mathbb{E}_W[W]\,\mathrm{Var}_X[X] \\ &= 0\end{aligned} \tag{144}$$

where for the last equality it has been used that

$$\begin{aligned}\mathbb{E}_W[W] &= \frac{1}{n}\sum_{i=1}^{n} w_i f_W(w_i) \\ &= \frac{1}{2}[(-1)\frac{1}{2} + 1\frac{1}{2}] \\ &= 0\end{aligned} \tag{145}$$

Example 3.17.

According to Definition 38 the variance is defined as positive definite. This means

$$\begin{aligned} 0 &\leq \operatorname{Var}_{X,Y}\left[\frac{X}{\sqrt{\operatorname{Var}_X[X]}} \pm \frac{Y}{\sqrt{\operatorname{Var}_Y[Y]}}\right] \\ &= \frac{\operatorname{Var}_X[X]}{\operatorname{Var}_X[X]} + \frac{\operatorname{Var}_Y[Y]}{\operatorname{Var}_Y[Y]} \pm \frac{2}{\sqrt{\operatorname{Var}_X[X]\operatorname{Var}_Y[Y]}} \operatorname{Cov}_{X,Y}[X,Y] \\ &= 2 \pm 2\operatorname{Corr}_{X,Y}[X,Y]. \end{aligned} \tag{146}$$

From equation (146) *the result follows*

$$-1 \leq \operatorname{Corr}_{X,Y}[X,Y] \leq 1. \tag{147}$$

Example 3.18.

Show that if $Y = aX + b$ *given parameters* $a > 0$ *and* b*, then* $\operatorname{Corr}_{X,Y}[X,Y] = 1$*. Similarly show that if* $a < 0$*, then* $\operatorname{Corr}_{X,Y}[X,Y] = -1$*.*

Since $Y = X^2$ *is uniquely determined by* X*,*

$$\operatorname{Corr}_{X,Y}[X,Y] = \operatorname{Corr}_X[X, aX+b]. \tag{148}$$

Using Definition 42 and Definition 41

$$\operatorname{Corr}_X[X, aX+b] = \frac{\operatorname{Cov}_X[X, aX+b]}{\sqrt{\operatorname{Var}_X[X]\operatorname{Var}_X[aX+b]}} \tag{149}$$

$$\begin{aligned} \operatorname{Cov}_X[X, aX+b] &= \mathbb{E}_X[X(aX+b)] - \mathbb{E}_X[X]\mathbb{E}_X[aX+b] \\ &= a\mathbb{E}_X[X^2] + b\mathbb{E}_X[X] - a\mathbb{E}_X[X]^2 - b\mathbb{E}_X[X] \\ &= a\operatorname{Var}_X[X] \end{aligned} \tag{150}$$

$$\begin{aligned} \operatorname{Var}_X[aX+b] &= a^2\operatorname{Var}_X[X] + \cancelto{0}{\operatorname{Var}_X[b]} + \cancelto{0}{2\operatorname{Cov}_{X,X}[aX, b]} \\ &= a^2\operatorname{Var}_X[X] \end{aligned} \tag{151}$$

$$\begin{aligned} \operatorname{Corr}_X[X, aX+b] &= \frac{a\operatorname{Var}_X[X]}{\sqrt{a^2\operatorname{Var}_X[X]\operatorname{Var}_X[X]}} \\ &= \frac{a}{|a|} \end{aligned} \tag{152}$$

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

CHAPTER 4

Assigning Probability Functions

While Chapter 3 provides the formal definition of probability measures and their manipulation, it is not sufficient on its own to conduct inference. In practice, one must also specify the probability measure or, equivalently, the probability density or mass functions. Assigning these functions requires a principled method to convert available information into a probability distribution.

The central challenge is to incorporate only the information that is actually known, without introducing unwarranted assumptions about unknown quantities. Logical analysis provides the framework for this task: it ensures that probability assignments are internally consistent and make full use of the information at hand. Several approaches implement this principle. Logical analysis can be applied directly to the sum and product rules to construct probability functions [7]; it can exploit group invariances inherent in the problem [8]; and it can ensure consistent marginalization of nuisance parameters [9].

Among these methods, the principle of maximum entropy[14] stands out for its generality and power. By selecting the probability distribution that maximizes entropy subject to the known constraints, it provides a systematic, non-arbitrary means of assigning probabilities while remaining maximally noncommittal about unknown information [8, 10–13]. The remainder of this chapter develops the maximum entropy principle and illustrates how it can be applied to derive probability distributions from partial knowledge.

4.1 THE PRINCIPLE OF MAXIMUM ENTROPY

The principle of maximum entropy, first proposed by Jaynes [14], 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

$$Z : \Omega \to \Omega_Z \tag{153}$$

be a generic random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$, where $\mathbb{P}_Z$ is the image measure of $\mathbb{P}$. The goal of the maximum entropy principle is to determine the probability measure $\mathbb{P}_Z$ that best represents the current state of knowledge, given background information I and a set of moment constraints.

Definition 44 (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.

From definition Definition 39, the image measure $\mathbb{P}_Z$ admits a density f_Z with respect to the measure μ, so that for any event $B \in \mathcal{F}_Z$,

$$\mathbb{P}_Z(B \mid \gamma, I) = \int_B f_Z(z \mid \gamma, I) d\mu(z), \tag{154}$$

where f_Z is the PMF for a discrete sample space and the PDF for a continuous sample space. The maximum entropy principle asserts that the density $f_Z(z|\gamma, I)$ that best represents the current state of knowledge is the one that maximizes the constrained Shannon entropy [3], where γ denotes the parameters of the distribution and I the background information. The Shannon entropy of a probability density f_Z can be written

$$H[f_Z] = -\int_{\Omega_Z} f_Z(z \mid \gamma, I) \ln \frac{f_Z(z \mid \gamma, I)}{m(z)} d\mu(z), \tag{155}$$

where $m(z)$ is a reference measure that ensures invariance of the entropy under reparameterizations of z. To incorporate known constraints, such as moment

conditions or normalization, one introduces Lagrange multipliers $\gamma_0, \gamma_1, \ldots, \gamma_n$ and defines the Lagrangian functional, which represents the constrained entropy:

$$\mathcal{L}[f_Z] = -\int_{\Omega_Z} f_Z(z \mid \gamma, I)\left(\ln \frac{f_Z(z \mid \gamma, I)}{m(z)} + \gamma_0 + \sum_{j=1}^{n} \gamma_j C_j(z)\right) d\mu(z), \quad (156)$$

where each $C_j(z)$ denotes a constraint function. Maximizing $\mathcal{L}[f_Z]$ with respect to f_Z yields the probability distribution of maximum entropy consistent with the given constraints. The maximum of $\mathcal{L}$ with respect to f_Z is defined by the Euler-Lagrange condition

$$\frac{\partial}{\partial f_Z} f_Z\Big(\ln \tfrac{f_Z}{m} + \gamma_0 + \sum_{j=1}^{n} \gamma_j C_j\Big) = 0, \quad (157)$$

which simplifies to

$$\ln \frac{f_Z(z \mid \gamma, I)}{m(z)} + 1 + \gamma_0 + \sum_{j=1}^{n} \gamma_j C_j(z) = 0. \quad (158)$$

Hence the maximum-entropy distribution takes the exponential family form

$$\begin{aligned} f_Z(z \mid \gamma, I) &= m(z) e^{-1-\gamma_0-\sum_{j=1}^{n} \gamma_j C_j(z)} \\ &= \frac{m(z) e^{-\sum_{j=1}^{n} \gamma_j C_j(z)}}{\int_{\Omega_Z} m(z') e^{-\sum_{j=1}^{n} \gamma_j C_j(z')} d\mu(z')}. \end{aligned} \quad (159)$$

The constants γ_j are determined by the imposed constraints. The resulting probability measure (Equation 154) defines the unique maximum-entropy probability measure consistent with the given information.

Example 4.1.

Consider a continuous random variable, Z, with sample space $\Omega_Z = \mathbb{R}$, assumed to be symmetric around a mean μ and with variance σ^2. In this case

$$f_Z(z \mid \gamma, I) = m(z)e^{-1-\gamma_0-\gamma_1 z-\gamma_2 z^2}. \tag{160}$$

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

$$\begin{aligned}\int_{\Omega_Z} f_Z(z|\gamma, I)dz &= me^{-1-\gamma_0}\int_{\Omega_Z} e^{-\gamma_1 z-\gamma_2 z^2}dz \\ &= me^{-1-\gamma_0}\sqrt{\frac{\pi}{\gamma_2}}e^{\frac{\gamma_1^2}{4\gamma_2}} \\ &= 1.\end{aligned} \tag{161}$$

Defining $K^{-1} = me^{-1-\gamma_0}$ yields

$$\begin{aligned}f_Z(z|\gamma, I) &= \frac{e^{-\gamma_1 z-\gamma_2 z^2}}{K} \\ &= \sqrt{\frac{\gamma_2}{\pi}}e^{-\frac{\gamma_1^2}{4\gamma_2}-\gamma_1 z-\gamma_2 z^2}.\end{aligned} \tag{162}$$

Now, imposing the mean constraint

$$\begin{aligned}\int_{\Omega_Z} z f_Z(z|\gamma, I)dz &= \frac{\int_{\Omega_Z} ze^{-\gamma_1 z-\gamma_2 z^2}dz}{K} \\ &= -\frac{\gamma_1}{2\gamma_2} \\ &= \mu.\end{aligned} \tag{163}$$

Hereby

$$\begin{aligned}f_Z(z|\gamma, I) &= \sqrt{\frac{\gamma_2}{\pi}}e^{-\mu^2\gamma_2+2\mu\gamma_2 z-\gamma_2 z^2} \\ &= \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{\mu-z}{\sigma}\right)^2},\end{aligned} \tag{164}$$

where $\sigma \equiv \frac{1}{2\gamma_2}$ has been defined. Equation 164 can be identified as the normal distribution.

Example 4.2.

Consider a continuous random variable, Z, with sample space $\Omega_S = [0,1]$. In order to impose the limited support, require that $\ln(z)$ and $\ln(1-z)$ be well defined. In this case

$$f_Z(z|\gamma, I) = m(z)e^{-1-\gamma_0-\gamma_1 \ln z-\gamma_2 \ln(1-z)}. \tag{165}$$

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

$$\begin{aligned}\int_{\Omega_Z} f_Z(z|\gamma, I) &= me^{-1-\gamma_0}\int_{\Omega_Z} z^{-\gamma_1}(1-z)^{-\gamma_2}dz \\ &= me^{-1-\gamma_0}\frac{\Gamma(1-\gamma_1)\Gamma(1-\gamma_2)}{\Gamma(2-\gamma_1-\gamma_2)} \\ &= 1.\end{aligned} \tag{166}$$

Now define $\alpha \equiv 1-\gamma_1 \wedge \beta \equiv 1-\gamma_2$. Hereby

$$f_Z(z|\alpha, \beta, I) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{167}$$

which is the beta distribution.

Example 4.3.

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

$$\begin{aligned}f_Z(z|\gamma, I) &= m(z)e^{-1-\gamma_0-\gamma_1 z-\gamma_2 \ln z} \\ &= \tilde{m}(z)z^{-\gamma_2}e^{-\gamma_1 z}\end{aligned} \tag{168}$$

where $\tilde{m}(z) = m(z)e^{-1-\gamma_0}$. Taking a uniform measure ($m(z) = const$) and imposing normalization

$$\begin{aligned}\int_{\Omega_Z} f_Z(z|\gamma, I)dz &= \tilde{m}\int \Omega_Z z^{-\gamma_2}e^{-\gamma_1 z}dz \\ &= 1.\end{aligned} \tag{169}$$

The integral is recognized as the Gamma function

$$\int_{\Omega_Z} z^{\alpha-1}e^{-\beta z}dz = \frac{\Gamma(\alpha)}{\beta^\alpha} \tag{170}$$

with $\alpha = 1-\gamma_2$ and $\beta = \gamma_1$. Substituting $\tilde{m}$, α, β back into Equation 168

$$f_Z(z|\gamma, I) = \frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}, \tag{171}$$

which is the Gamma distribution.

Example 4.4.

Consider a continuous random variable, Z, with sample space $\Omega_Z = [0, \infty)$ and known mean μ. In this case

$$f_Z(z|\gamma, I) = m(z)e^{-1-\gamma_0-\gamma_1 z}. \tag{172}$$

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

$$\begin{aligned} \int_{\Omega_Z} f_Z(z|\gamma, I)dz &= me^{-1-\gamma_0} \int_{\Omega_Z} e^{-\gamma_1 z}dz \\ &= me^{-1-\gamma_0} \frac{1}{\gamma_1} \\ &= 1 \end{aligned} \tag{173}$$

and the mean constraint

$$\begin{aligned} \int_{\Omega_Z} zp(z|\gamma, I)dz &= \int_{\Omega_Z} z\gamma_1 e^{-\gamma_1 z}dz \\ &= \frac{1}{\gamma_1} \\ &= \mu. \end{aligned} \tag{174}$$

Combining Equation 173, Equation 174 and Equation 172 yields

$$f_Z(z|\gamma, I) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}, \tag{175}$$

which is the Gamma distribution.

Example 4.5.

Consider a discrete random variable, Z, with sample space $\Omega_Z = \{0, 1\}$ and mean μ. In this case

$$f_Z(z|\gamma, I) = m(z)e^{-1-\gamma_0-\gamma_1 z}. \tag{176}$$

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

$$\begin{aligned} \sum_{z=0}^{1} f_Z(z) &= me^{-1-\gamma_0} \left(1 + e^{-\gamma_1}\right) \\ &= 1 \end{aligned} \tag{177}$$

and mean constraint

$$\begin{aligned}\sum_{z=0}^{1} z f_Z(z) &= m e^{-1-\gamma_0} e^{-\gamma_1} \\ &= \frac{1}{1+e^{\gamma_1}} \\ &= \mu\end{aligned} \tag{178}$$

This means

$$\begin{aligned}f_Z(0|\gamma, I) &= m e^{-1-\gamma_0} \\ &= \frac{1}{1+e^{-\gamma_1}} \\ &= 1-\mu\end{aligned} \tag{179}$$

and

$$\begin{aligned}f_Z(1|\gamma, I) &= m e^{-1-\gamma_0-\gamma_1} \\ &= \mu,\end{aligned} \tag{180}$$

or

$$f_Z(z|\gamma, I) = \mu^z (1-\mu)^{1-z}. \tag{181}$$

which is the Bernoulli distribution.

Example 4.6.

Consider a discrete random variable, Z, with sample space $\Omega_Z = \{0, 1, \ldots, n\}$ representing the total number of successes in n independent Bernoulli trials with mean μ. In this case

$$f_Z(z|\gamma, I) = m(z) e^{-\gamma_0 - \gamma_1 z}. \tag{182}$$

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

$$\begin{aligned}\sum_{z=0}^{n} f_Z(z|\gamma, I) &= \sum_{z=0}^{n} \binom{n}{z} e^{-\gamma_0 - \gamma_1 z} \\ &= 1,\end{aligned} \tag{183}$$

yields

$$e^{-\gamma_0} = (1 + e^{-\gamma_1})^{-n}. \tag{184}$$

The mean constraint

$$\begin{aligned}\sum_{z=0}^{n} z f_Z(z|\gamma, I) &= n\frac{e^{-\gamma_1}}{1+e^{-\gamma_1}} \\ &= n\mu\end{aligned} \tag{185}$$

gives

$$e^{-\gamma_1} = \frac{\mu}{1-\mu}. \tag{186}$$

Finally, substituting $e^{-\gamma_0}$ *and* $e^{-\gamma_1}$ *into* $f_Z(z|\gamma, I)$ *gives the maximum entropy distribution*

$$f_Z(z|\gamma, I) = \binom{n}{z}\mu^z(1-\mu)^{n-z}, \tag{187}$$

which is the Binomial distribution.

Example 4.7.

Consider a discrete random variable Z *with sample space* $\Omega_Z = \mathbb{N}_0$ *with a known mean* μ*. In this case*

$$f_Z(z|\gamma, I) = m(z)e^{-1-\gamma_0-\gamma_1 z}. \tag{188}$$

Take the counting measure $m(z) = 1/z!$ *and impose the normalization constraint*

$$\begin{aligned}\sum_{z=0}^{\infty} f_Z(z|\gamma, I) &= \sum_{z=0}^{\infty} \frac{e^{-1-\gamma_0-\gamma_1 z}}{z!} \\ &= e^{-1-\gamma_0}\sum_{z=0}^{\infty}\frac{e^{-\gamma_1 z}}{z!} \\ &= 1,\end{aligned} \tag{189}$$

Identifying the sum with the Taylor expansion

$$\sum_{z=0}^{\infty}\frac{e^{-\gamma_1 z}}{z!} = e^{e^{-\gamma}} \tag{190}$$

yields

$$e^{-1-\gamma_0} = e^{-e^{-\gamma_1}} \quad \Rightarrow \quad 1+\gamma_0 = e^{-\gamma_1}. \tag{191}$$

Imposing the mean constraint

$$\begin{aligned}\sum_{z=0}^{\infty} z f_Z(z|\gamma, I) &= e^{-1-\gamma_0} \sum_{z=1}^{\infty} \frac{z e^{-\gamma_1 z}}{z!} \\ &= e^{-1-\gamma_0} \sum_{z=1}^{\infty} \frac{e^{-\gamma_1 z}}{(z-1)!} \\ &= e^{-\gamma_1} e^{-1-\gamma_0} \sum_{y=0}^{\infty} \frac{e^{-\gamma_1 y}}{y!} \\ &= e^{-\gamma_1} \\ &= \mu\end{aligned} \tag{192}$$

where Equation 189 and $y = z - 1$ *has been used. Combining Equation 188, Equation 191 and Equation 192 yield*

$$f_Z(z|\gamma, I) = \frac{\mu^z e^{-\mu}}{z!}. \tag{193}$$

which is the Poisson distribution.

CHAPTER 5

Introduction to Statistics

Let the observed outcome of a statistical experiment be described by the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Unlike in probability theory, where $\mathbb{P}$ is assumed known, in statistics the data-generating measure is typically unknown and must be inferred from observations. Let [4, 5, 15, 16]

$$X_i : \Omega \to \Omega_{X_i} \tag{194}$$

be a generic random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\Omega_{X_i}, \mathcal{F}_{X_i})$, with the image measure

$$\mathbb{P}_{X_i} = \mathbb{P} \circ X_i^{-1}. \tag{195}$$

The joint image measure of n such random variables,

$$\mathbb{P}_{X_1,\ldots,X_n} = \mathbb{P} \circ (X_1, \ldots, X_n)^{-1}, \tag{196}$$

is defined on the product measurable space

$$\big(\Omega_{X_1} \times \cdots \times \Omega_{X_n}, \mathcal{F}_{X_1} \otimes \cdots \otimes \mathcal{F}_{X_n}\big), \tag{197}$$

which for brevity will be written as $\mathbb{P}_{X_{1:n}}$ and $(\Omega_{X_{1:n}}, \mathcal{F}_{X_{1:n}})$, respectively.

Definition 45 (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 46 (Parametric Family of Probability Measures)**.** *A parametric family (or parametric subset) of probability measures is a set of the form*

$$\mathcal{P}' = \{\mathbb{P}^w_{X_{1:n}} \mid w \in \Omega_W\} \subseteq \mathcal{P}, \tag{198}$$

where Ω_W is the parameter space. For each fixed w, $\mathbb{P}^w_{X_{1:n}}$ is a probability measure on $(\Omega_{X_{1:n}}, \mathcal{F}_{X_{1:n}})$.

Definition 47 (Parameter Space)**.** *The parameter space Ω_W is the set of all values w that index the probability measures $\mathbb{P}^w_{X_{1:n}} \in \mathcal{P}'$.*

Definition 48 (Identifiable Statistical Model)**.** *A parametric family* $\mathcal{P}' = \{\mathbb{P}^w_{X_{1:n}} \mid w \in \Omega_W\}$ *is identifiable if the mapping*

$$w \in \Omega_W \mapsto \mathbb{P}^w_{X_{1:n}} \in \mathcal{P}' \tag{199}$$

is injective; i.e., distinct parameter values induce distinct probability measures.

The parameters $w \in \Omega_W$ may be interpreted in two different ways in practice: either as fixed but unknown constants or as realizations of a random variable.

Axiom 3 (Parameter Fixedness)**.** *The parameter* $w \in \Omega_W$ *is treated as a fixed but unknown constant. In this setting the image measure* $\mathbb{P}_{X_{1:n}}$ *is assumed to equal* $\mathbb{P}^{w^*}_{X_{1:n}}$ *for some (unknown)* $w^* \in \Omega_W$*.*

Axiom 4 (Parameter as a Random Variable)**.** *The parameter* $w \in \Omega_W$ *is treated as the realization of a random variable*

$$W : \Omega \to \Omega_W \tag{200}$$

from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_W, \mathcal{F}_W)$*, with the image (prior) measure*

$$\mathbb{P}_W = \mathbb{P} \circ W^{-1}. \tag{201}$$

The joint image measure $\mathbb{P}_{W,X_{1:n}}$ *is defined on*

$$(\Omega_W \times \Omega_{X_{1:n}}, \mathcal{F}_W \otimes \mathcal{F}_{X_{1:n}}). \tag{202}$$

Remark 16 (Parameter interpretation)**.** *For both Axiom 3 and Axiom 4, the value of a parameter is considered fixed. Axiom 4 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.*

5.1 INTERPRETATION OF PROBABILITY MEASURES

Although probability measures are defined according to Definition 30, their interpretation is not defined beyond this definition. For this reason there are two broadly accepted interpretations of probability; objective and subjective.

Definition 49 (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 [17]*

$$\mathbb{P}(E) \equiv \lim_{n\to\infty} \left(\frac{m}{n}\right) \tag{203}$$

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

Definition 50 (Sugeno Measure)**.** *Let* $(\Omega, \mathcal{F})$ *be a measurable space. A set function*

$$\text{Bel} : \mathcal{F} \to [0, 1] \tag{204}$$

is called a Sugeno measure [18] if it satisfies the following properties:

1. *Axiom 1 (non-negativity),*
2. $\text{Bel}(\Omega) = 1$ *(normalization),*
3. $\text{Bel}(A) \leq \text{Bel}(B)$ *for all* $A, B \in \mathcal{F}$ *with* $A \subseteq B$ *(monotonicity).*

Definition 51 (Subjective Probability Measure)**.** *A subjective probability measure is a numerical representation of rational beliefs. Formally, it is a probability measure* $\mathbb{P}$*, according to Definition 30, on a measurable space* $(\Omega, \mathcal{F})$ *that fulfills Definition 50 [18, 19].*

Theorem 11 (Probability vs. Sugeno Measures)**.** *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. □

Remark 17. *Since a probability measure $\mathbb{P}$ satisfies the axioms of a Sugeno measure, it can be interpreted as a belief function.*

Definition 52 (Frequentist Statistics)**.** *Frequentist statistics is a paradigm that adopts Axiom 3 and Definition 49 of probability.*

Definition 53 (Bayesian Statistics)**.** *Bayesian statistics is a paradigm that adopts Axiom 4 and definition Definition 51 of probability.*

Remark 18 (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 in which the observed data are generated from a combination of a Normal distribution with parameters μ, σ and a Beta distribution with parameters a, b. Define the parameter random vector

$$W = \begin{pmatrix} W_\mu \\ W_\sigma \\ W_a \\ W_b \end{pmatrix} : \Omega \to \Omega_W, \tag{205}$$

where each component $W_\mu, W_\sigma, W_a, W_b$ is a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Omega_i, \mathcal{F}_i)$ with an image (prior) measure $\mathbb{P}_i$ $\forall i \in \{\mu, \sigma, a, b\}$.

5.2 FRAMING OF STATISTICS

In this book, statistics is framed as a game against Nature, following conventions from decision theory[2]. In this game, there are two players, whose roles are formalized in Definition 54 and Definition 55.

Definition 54 (Robot)**.** *The Robot is the primary decision maker in the statistical game.*

Definition 55 (Nature)**.** *Nature is an unpredictable decision maker that can interfere with the Robot's outcomes. It models uncertainty in the decision-making process.*

Remark 19 (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 56 (Cost Function)**.** *A cost function associates a numerical penalty depending on decision* $u \in \Omega_U$ *and* $s \in \Omega_S$*,*

$$C : \Omega_U \times \Omega_S \to \mathbb{R}. \tag{206}$$

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

$$D = \{(X = x_i, S = s_i)\}_{i=1}^{n}, \tag{207}$$

the Robot's objective is to formulate a decision rule that minimize the expected cost associated with its decisions[1].

Definition 57 (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*

$$U : \Omega_X \times (\Omega_X \times \Omega_S)^n \to \Omega_U, \tag{208}$$

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.

$$\Omega_U = \{\text{"}bring\ umbrella\text{"}, \text{"}don't\ bring\ umbrella\text{"}\}. \tag{209}$$

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

$$\Omega_S = \{\text{"}rain\text{"}, \text{"}no\ rain\text{"}\}, \tag{210}$$

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 $X : \Omega \to \Omega_X$ represent the information available (the information may be missing or null) to the Robot regarding the decision Nature will make, while $S : \Omega \to \Omega_S$ 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.

Remark 20 (Relaxation of Notation)**.** *The formal measure-theoretic details introduced so far provide the foundation for statistical reasoning. In practice, however, many of these technicalities can be safely abstracted to facilitate computations and exposition. Accordingly, in the remainder of this book, the notation around probability spaces, σ-algebras, and probability measures will be* relaxed*. Specifically:*

- *The symbol p will be used informally to denote probability distributions, densities, mass functions, or measures.*
- *The probability of a random variable taking a specific value, $p(X = x)$, will usually be written as $p(x)$ for brevity.*

This simplification allows advanced manipulation of probabilities without cumbersome formalism. Nevertheless, familiarity with the formal definitions provided here remains beneficial for a rigorous understanding.

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[1]

$$\begin{aligned}\mathbb{E}[C(U,S)|I] &= \int dDdxdsC(U(x,D),s)p(X=x,S=s,D|I) \\ &= \int d\tilde{D}dsC(U(\tilde{D}),s)p(S=s,\tilde{D}|I)\end{aligned} \tag{211}$$

where $\tilde{D} = \{D, X = x\}$, I denotes the background information (Definition 44) and the Robot aims to find the decision rule (Definition 57) which minimizes Equation 211, meaning

$$U^* = \arg\min_U \mathbb{E}_{\tilde{D}}[C(U,S)|I]. \tag{212}$$

From Theorem 7

$$\mathbb{E}_{\tilde{D}}[C(U,S)|I] = \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]]. \tag{213}$$

Using Equation 213 in Equation 212

$$\begin{aligned}U^* &= \arg\min_U \mathbb{E}_{\tilde{D}}[\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I]] \\ &= \arg\min_U \int d\tilde{D}p(\tilde{D}|I)\mathbb{E}_{S|\tilde{D}}[C(U,S)|\tilde{D},I].\end{aligned} \tag{214}$$

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

$$\begin{aligned}U^*(\tilde{D}) &= \arg\min_{U(\tilde{D})} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] \\ &= \arg\min_{U(\tilde{D})} \int dsC(U(\tilde{D}),s)p(s|\tilde{D},I).\end{aligned} \tag{215}$$

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

$$\begin{aligned}\tilde{D} &= \{S = s_i\}_{i=1}^n \\ &\equiv D_s.\end{aligned} \tag{216}$$

In this case, the Robot is forced to model the decisions of Nature with a probability distribution with associated parameters without observations. From Equation 215 the optimal action for the Robot can be written

$$U^*(D_s) = \arg\min_{U(D_s)} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}),S)|\tilde{D},I] \tag{217}$$

5.2.1 *Assigning a Cost Function*

The cost function (see definition 56) 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 [2], 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 58 (Linear Cost Function)**.** *The linear cost function is defined viz*

$$C(U(\tilde{D}), s) \equiv |U(\tilde{D}) - s|. \tag{218}$$

Theorem 12 (Median Decision Rule)**.** *Assuming the cost function of Definition 58*

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \int_{-\infty}^{\infty} ds|U(\tilde{D}) - s|p(s|\tilde{D}, I) \\ &= \int_{-\infty}^{U(\tilde{D})} ds(s - U(\tilde{D}))p(s|\tilde{D}, I) \\ &\quad + \int_{U(\tilde{D})}^{\infty} ds(U(\tilde{D}) - s)p(s|\tilde{D}, I)\end{aligned} \tag{219}$$

$$\begin{aligned}0 &= \frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}, I) + \int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D}, I) \\ &\quad + (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}, I) - \int_{U^*(\tilde{D})}^{\infty} dsp(s|\tilde{D}, I)\end{aligned} \tag{220}$$

$$\begin{aligned}\int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D}, I) &= \int_{U^*(\tilde{D})}^{\infty} dsp(s|\tilde{D}, I) \\ &= 1 - \int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D}, I)\end{aligned} \tag{221}$$

$$\int_{-\infty}^{U^*(\tilde{D})} dsp(s|\tilde{D}, I) = \frac{1}{2} \tag{222}$$

which is the definition of the median.

Definition 59 (Quadratic Cost Function)**.** *The quadratic cost function is defined as*

$$C(U(\tilde{D}), s) \equiv (U(\tilde{D}) - s)^2. \tag{223}$$

Theorem 13 (Expectation Decision Rule)**.** *Assuming the cost function of Definition 59*

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \int ds(U(\tilde{D}) - s)^2 p(s|\tilde{D}, I) \\ &\Downarrow \\ \frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(x)} &= 2U^*(\tilde{D}) - 2\int dssp(s|\tilde{D}, I) \\ &= 0 \\ &\Downarrow \\ U^*(\tilde{D}) &= \int dssp(s|\tilde{D}, I) \\ &= \mathbb{E}_{S|\tilde{D}}[S|\tilde{D}, I] \end{aligned} \tag{224}$$

which is the definition of the expectation value.

Definition 60 (0-1 Cost Function)**.** *The 0-1 cost function is defined viz*

$$C(U(\tilde{D}), s) \equiv 1 - \delta(U(\tilde{D}) - s). \tag{225}$$

Theorem 14 (MAP Decision Rule)**.** *The maximum aposteriori (MAP) follows from assuming 0-1 loss viz*

$$\mathbb{E}_{S|\tilde{D}}[C((\tilde{D}), S)|\tilde{D}, I] = 1 - \int ds\delta(U(\tilde{D}) - s)p(S = s|\tilde{D}, I) \tag{226}$$

meaning

$$\begin{aligned} \frac{\partial \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} &= -\frac{\partial p(S = U(\tilde{D})|\tilde{D}, I)}{\partial U(\tilde{D})}\Big|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= 0 \end{aligned} \tag{227}$$

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

$$C(U(\tilde{D}), s) = \alpha \cdot \text{swish}(U(\tilde{D}) - s, \beta) + (1 - \alpha) \cdot \text{swish}(s - U(\tilde{D}), \beta), \quad (228)$$

where

$$\text{swish}(z, \beta) = \frac{z}{1 + e^{-\beta z}}. \quad (229)$$

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

$$\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] = \int dsp(S = s|\tilde{D}, I)\, C(U(\tilde{D}), s). \quad (230)$$

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

$$\begin{aligned} \frac{dC}{dU} &= \frac{dC}{dz}\frac{dz}{dU} \\ &= \left(\frac{\alpha}{1 + e^{-\beta z}} - \frac{1 - \alpha}{1 + e^{\beta z}} \right. \\ &\qquad \left. + \frac{\alpha\beta e^{-\beta z} z}{(1 + e^{-\beta z})^2} + \frac{(1 - \alpha)\beta e^{\beta z} z}{(1 + e^{\beta z})^2} \right) \frac{dz}{dU} \\ &= \frac{\beta z e^{\beta z} - e^{\beta z} - 1}{(1 + e^{\beta z})^2} + \alpha + \mathcal{O}(\alpha^2) \\ &\approx \alpha - \frac{1}{(1 + e^{\beta z})^2} \end{aligned} \quad (231)$$

leading to the approximate expected cost

$$\begin{aligned} \frac{d\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I]}{dU(\tilde{D})} &\approx \int dsp(s|\tilde{D}, I) \left(\alpha - \frac{1}{(1 + e^{\beta z(\tilde{D}, s)})^2} \right) \\ &= \alpha - \int dsp(s|\tilde{D}, I) \frac{1}{(1 + e^{\beta z(\tilde{D}, s)})^2}. \\ &= 0 \end{aligned} \quad (232)$$

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,

$$\int_{-\infty}^{\infty} dsp(s|\tilde{D}, I) \frac{1}{(1 + e^{\beta z(\tilde{D}, s)})^2} \approx \int_{U(\tilde{D})}^{\infty} dsp(s|\tilde{D}, I) \quad (233)$$

This means the optimal decision rule can be written viz

$$\alpha \approx \int_{U(\tilde{D})}^{\infty} dsp(s|\tilde{D}, I). \tag{234}$$

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

$$\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] = \sum_{s \in \Omega_S} C(U(\tilde{D}), s)p(s|\tilde{D}, I), \tag{235}$$

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

$$D = \{(x_1, s_1), (x_2, s_2), (x_3, s_3)\} \tag{236}$$

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

$$D = \{(x_i, s_i)\}_{i=1}^{n} \tag{237}$$

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

$$U_k(x, D) = \begin{cases} u^{(1)}, & p(S = s^{(2)}|x, D, I) < k, \\ u^{(2)}, & p(S = s^{(2)}|x, D, I) \geq k. \end{cases} \tag{238}$$

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

$$\mathrm{TPR(k)} = \frac{\mathrm{TP(k)}}{\mathrm{TP(k)} + \mathrm{FN(k)}}, \tag{239}$$

$$\mathrm{FPR(k)} = \frac{\mathrm{FP(k)}}{\mathrm{FP(k)} + \mathrm{TN(k)}}, \tag{240}$$

$$\mathrm{Accuracy(k)} = \frac{\mathrm{TP(k)} + \mathrm{TN(k)}}{\mathrm{TP(k)} + \mathrm{TN(k)} + \mathrm{FP(k)} + \mathrm{FN(k)}}. \tag{241}$$

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

$$\mathrm{ROC} = \{(\mathrm{FPR(k)}, \mathrm{TPR(k)}) : k \in [0, 1]\}. \tag{242}$$

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

$$\mathrm{AUROC} = p(p(S = s^{(2)}|X_{(s^{(2)})}, D, I) > p(S = s^{(2)}|X_{(s^{(1)})}, D, I)|D, I), \tag{243}$$

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,

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(u))\, du, \tag{244}$$

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

$$\mathrm{AR} = 2 \cdot \mathrm{AUROC} - 1. \tag{245}$$

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 235)

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \sum_s C(U(\tilde{D}), S = s) p(S = s|\tilde{D}, I) \\ &= C(U(\tilde{D}), S = s^{(1)}) p(S = s^{(1)}|\tilde{D}, I) \\ &\quad + C(U(\tilde{D}), S = s^{(2)}) p(S = s^{(2)}|\tilde{D}, I) \end{aligned} \tag{246}$$

   For the different possible actions

$$\begin{aligned} \mathbb{E}_{S|\tilde{D}}[C(u^{(1)}, S)|\tilde{D}, I] &= \lambda_{12} p(S = s^{(2)}|\tilde{D}, I), \\ \mathbb{E}_{S|\tilde{D}}[C(u^{(2)}, S)|\tilde{D}, I] &= \lambda_{21} p(S = s^{(1)}|\tilde{D}, I), \end{aligned} \tag{247}$$

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

$$\mathbb{E}_{S|\tilde{D}}[C(u^{(1)}, S)|\tilde{D}, I] < \mathbb{E}_{S|\tilde{D}}[C(u^{(1)}, S)|\tilde{D}, I]) \tag{248}$$

meaning

$$\begin{aligned}\lambda_{12}p(S=s^{(2)}|\tilde{D},I) &< \lambda_{21}p(S=s^{(1)}|\tilde{D},I)\\ &= \lambda_{21}(1-p(S=s^{(2)}|\tilde{D},I))\end{aligned} \tag{249}$$

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

$$p(S=s^{(2)}|\tilde{D},I) < \frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} = k \tag{250}$$

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*

$$C(U(\tilde{D}),s)=\begin{cases}0 & \text{if correct classification } (U(\tilde{D})=s)\\ \lambda_r & \text{if reject option } (U(\tilde{D})=\text{reject})\\ \lambda_s & \text{if wrong classification } (U(\tilde{D})\neq s)\end{cases}. \tag{251}$$

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

$$p(S=U(\tilde{D})|\tilde{D},I) \geq p(S\neq U(\tilde{D})|\tilde{D},I) \tag{252}$$

and if

$$p(S=U(\tilde{D})|\tilde{D},I) \geq 1-\frac{\lambda_r}{\lambda_s}. \tag{253}$$

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

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] &= \sum_s C(U(\tilde{D}), S = s)p(S = s|\tilde{D}, I) \\ &= \sum_{s \neq U(\tilde{D})} \lambda_s p(S = s|\tilde{D}, I) \\ &= \lambda_s(1 - p(S = U(\tilde{D})|\tilde{D}, I))\end{aligned} \tag{254}$$

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*

$$\begin{aligned}\mathbb{E}_{S|\tilde{D}}[C(\text{reject}, S)|\tilde{D}, I] &= \lambda_r \sum_s p(S = s|\tilde{D}, I) \\ &= \lambda_r.\end{aligned} \tag{255}$$

Equation (254) *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 254 and Equation 255 yields*

$$(1 - p(S = U(\tilde{D})|\tilde{D}, I))\lambda_s < \lambda_r \tag{256}$$

meaning

$$p(S = U(\tilde{D})|\tilde{D}, I) \geq 1 - \frac{\lambda_r}{\lambda_s}. \tag{257}$$

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

$$\frac{\lambda_r}{\lambda_s} = 0 \tag{258}$$

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

$$p(S = U(\tilde{D})|\tilde{D}, I) = 1, \tag{259}$$

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.

$$\frac{\lambda_r}{\lambda_s} = 1 \tag{260}$$

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 21 (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 53) and Frequentist (Definition 52) 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 53, which follows the definition of subjective probability (Definition 51) and the treating the parameters as realizations of a random variable (Axiom 4). The Bayesian framework originally come from the work of Bayes [20] and Laplace [21] with much of the modern discussions and formalism created later by Finetti [22] and Jeffreys [23] and Savage [24].

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 215 depend on the parameters $w_1, \dots w_n$ of the statistical model. Introducing the shorthand notation $W = w_1 \dots W = w_n \to w$, $dw_1 \dots dw_n \to dw$ and $X = x \to x$, then

$$\begin{aligned} p(s|x, D, I) &= \int dw p(w, s|x, D, I) \\ &= \int dw p(s|w, x, D, I) p(w|x, D, I) \end{aligned} \tag{261}$$

Example 5.9.

Writing out the shorthand notation

$$\begin{aligned} p(W = w_1, \dots, W = w_n, S = s|X = x, D, I) \to p(w, s|x, D, I), \\ dw_1 \dots dw_n \to dw. \end{aligned} \tag{262}$$

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

$$\begin{aligned} p(w|x, D, I) &= p(w|D, I) \\ &= \frac{p(D_s|w, D_x, I)p(w|I)}{p(D_s|D_x, I)}, \end{aligned} \tag{263}$$

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 5 has been used for the first and second equality.

Axiom 5 (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 44). From Equation 263, 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,

$$f : \Omega_W \times \Omega_X \to \mathbb{R}, \tag{264}$$

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 59 will be considered to review the subject. From Theorem 13 the best action for the Robot can be written

$$U^*(x, D) = \int dssp(s|x, D, I) \tag{265}$$

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

$$p(s|x, w, \xi, I) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(w,x)-s)^2}. \tag{266}$$

Using Equation 266 and marginalizing over ξ, w

$$\begin{aligned} p(s|x, D, I) &= \int dwd\xi p(s, w, \xi|x, D, I) \\ &= \int dwd\xi p(s|x, w, \xi, D, I)p(w, \xi|x, D, I) \\ &= \int dwd\xi p(s|x, w, \xi, I)p(w, \xi|D, I), \end{aligned} \tag{267}$$

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 266) and $p(w, \xi|x, D, I) = p(w, \xi|D, I)$ from Axiom 5. Using Equation 267 in Equation 265[1]

$$\begin{aligned} U^*(x, D) &= \int dwd\xi f(w, x)p(w, \xi|D, I), \\ &= \mathbb{E}[f|x, D, I] \end{aligned} \tag{268}$$

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

$$\begin{aligned}\mathbb{E}[S|x,w,\xi,I] &= \int dssp(s|x,w,\xi,I) \\ &= f(w,x)\end{aligned} \tag{269}$$

according to Equation 266. Using Bayes theorem (Theorem 2)

$$p(w,\xi|D,I) = \frac{p(D_s|D_x,w,\xi,I)p(w,\xi|D_x,I)}{p(D_s|D_x,I)} \tag{270}$$

where from marginalization (Theorem 3)

$$p(D_s|D_x,I) = \int dwd\xi p(D_s|D_x,w,\xi,I)p(w,\xi|D_x,I). \tag{271}$$

Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written (using equation Equation 266)

$$p(D_s|D_x,w,\xi,I) = \left(\frac{\xi}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{\xi}{2}(f(w,x_i)-s_i)^2} \tag{272}$$

From the chain rule (see Theorem 1) and Theorem 5

$$p(w,\xi|D_x,I) = p(w|\xi,I)p(\xi|I). \tag{273}$$

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.

$$\begin{aligned}p(w|\xi,I) &= p(w|I) \\ &= \int d\lambda p(w|\lambda,I)p(\lambda|I)\end{aligned} \tag{274}$$

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

$$p(w|\lambda,I)p(\lambda|I) = \prod_{q=1}^{\tilde{n}} \frac{\lambda_q^{\frac{n_q}{2}}}{(2\pi)^{\frac{n_q}{2}}} e^{-\frac{\lambda_q}{2}\sum_{l=1}^{n_q} w_l^2} \frac{\beta_q^{\alpha_q}}{\Gamma(\alpha_q)} \lambda_q^{\alpha_q-1} e^{-\beta_q\lambda_q} \tag{275}$$

where α_q, β_q are prior parameters (a part of the background information, Definition 44) and $\tilde{n}$ is the number of hyper parameters. In the completely general

2 The normally distributed prior is closely related to weight decay [25], 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

$$\begin{aligned} p(\xi|I) &= \mathrm{Ga}(\xi|\tilde{\alpha}, \tilde{\beta}) \\ &= \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \xi^{\tilde{\alpha}-1} e^{-\tilde{\beta}\xi}. \end{aligned} \tag{276}$$

At this point equation Equation 265 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 44) and can be approximated by obtaining samples from $p(w, \xi, \lambda|D, I)$ via Hamiltonian Monte Carlo (HMC) [26–29] (see Appendix A for a review of HMC). The centerpiece in the HMC algorithm is the Hamiltonian defined viz [28, 29]

$$H \equiv \sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \ln[p(w, \xi, \lambda|D, I)] + const, \tag{277}$$

where

$$p(w, \xi|D, I) = \int d\lambda p(w, \xi, \lambda|D, I). \tag{278}$$

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 [30–32]). Using Equation 270-Equation 278 yields

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} + \frac{n}{2}[\ln(2\pi) - \ln(\xi)] + \frac{\xi}{2} \sum_{i=1}^{n} (f(w, x_i) - s_i)^2 \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln(\lambda_q) + \beta_q \lambda_q \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} w_l^2 \Bigg) \\ &+ \ln(\Gamma(\tilde{\alpha})) - \tilde{\alpha} \ln(\tilde{\beta}) + (1 - \tilde{\alpha}) \ln(\xi) + \tilde{\beta}\xi + const. \end{aligned} \tag{279}$$

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 268 means $p(\theta, \xi, \lambda|D, I)$ is multiplied with ξ. Hence, when taking $-\ln(p(\theta, \xi, \lambda|D, I))$ according to Equation 277, a $-\ln(\xi)$ is added to the Hamiltonian. In practice this means

$$(1 - \tilde{\alpha}) \ln(\xi) \in H \Rightarrow -\tilde{\alpha} \ln(\xi). \tag{280}$$

5.3.2 *Bayesian Classification*

Classification involves the Robot building a model,

$$f : \Omega_W \times \Omega_X \to \Delta^K, \tag{281}$$

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

$$\Delta^K = \{p \in \mathbb{R}^K \mid p_s \geq 0, \sum_{s=1}^{K} p_s = 1\} \tag{282}$$

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

$$p(S = s \mid x, w, I) = f_{S=s}(w, x), \tag{283}$$

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

$$\sum_{s \in \Omega_S} p(S = s \mid x, w, I) = 1. \tag{284}$$

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 60), with the addition of a reject option at cost λ. This means

$$C(U(\tilde{D}), s) = 1 - \delta_{U(\tilde{D}),s} + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}}. \tag{285}$$

The optimal decision rule for the robot can the be written (Equation 235)

$$\begin{aligned} U^*(\tilde{D}) &= \arg\min_{U(\tilde{D})} \mathbb{E}_{S|\tilde{D}}[C(U(\tilde{D}), S)|\tilde{D}, I] \\ &= \arg\min_{U(\tilde{D})} \left(\sum_{s \in \Omega_S} C(U(\tilde{D}), s) p(S = s|\tilde{D}, I) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right) \\ &= \arg\min_{U(\tilde{D})} \left(1 - p(S = U(\tilde{D})|\tilde{D}, I) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right). \end{aligned} \tag{286}$$

In absence of the reject option, the optimal decision rule is to pick the MAP, similar to Theorem 14. Using Equation 283 and marginalizing over w

$$\begin{aligned} p(S = U(\tilde{D})|\tilde{D}, I) &= \int dw p(S = U(\tilde{D}), w|\tilde{D}, I) \\ &= \int dw p(S = U(\tilde{D})|w, \tilde{D}, I) p(w|\tilde{D}, I) \\ &= \int dw p(S = U(\tilde{D})|x, w, I) p(w|D, I) \\ &= \int dw f_{S=U(\tilde{D})}(w, x) p(w|D, I) \\ &= \mathbb{E}[f_{S=U(\tilde{D})}(w, x)|D, I], \end{aligned} \tag{287}$$

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

$$p(S = U(\tilde{D})|w, \tilde{D}, I) = p(S = U(\tilde{D})|w, x, I) \tag{288}$$

since by definition f (see Equation 283) produce a $1 - 1$ map of the input x and $p(w|\tilde{D}, I) = p(w|D, I)$ from Axiom 5. From Bayes theorem

$$p(w|D, I) = \frac{p(D_s|D_x, w, I) p(w|D_x, I)}{p(D_s|D_x, I)}, \tag{289}$$

where from Axiom 5 $p(w|D_x, I) = p(w|I)$. Assuming the distribution over w is normally distributed with zero mean and a precision described by a hyperparameter, λ,

$$p(w|I) = \int d\lambda p(w|\lambda, I) p(\lambda|I). \tag{290}$$

where $p(w|\lambda, I)p(\lambda|I)$ is given by Equation 275. Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written [33]

$$\begin{aligned} p(D_s|D_x, w, I) &= \prod_{i=1}^{n} p(S = s_i|X = x_i, w, I) \\ &= \prod_{i=1}^{n} f_{S=s_i}(w, x_i) \end{aligned} \quad . \tag{291}$$

At this point Equation 286 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

$$H \equiv \sum_q \sum_l \frac{p_l^2}{2m_l} - \ln(p(w, \lambda|D, I)) + const \tag{292}$$

where

$$p(w|D, I) = \int d\lambda p(w, \lambda|D, I). \tag{293}$$

Using Equation 287-Equation 291 in equation (292) yields the Hamiltonian

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \sum_{i=1}^{n} \ln(f_{S=s_i}(w, x_i)) + \text{const} \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln(\Gamma(\alpha_q)) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln(\lambda_q) + \beta_q \lambda_q. \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln(\lambda_q)) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} w_l^2 \Bigg) \end{aligned} \tag{294}$$

Sampling Equation 294 yields a set of coefficients which can be used to compute $\mathbb{E}[f_s(w, x)|D, I]$ which in turn (see Equation 286) 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 5 is disregarded in this context. From Equation 215

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{S|D}[C(U(D), S)|D, I] \tag{295}$$

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 295

$$\begin{aligned}\mathbb{E}[C(u_1, S)|D, I] &= \sum_{s=s_1,s_2} C(u_1, s)p(S = s|D, I),\\ \mathbb{E}[C(u_2, S)|D, I] &= \sum_{s=s_1,s_2} C(u_2, s)p(S = s|D, I),\end{aligned} \tag{296}$$

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

$$\frac{p(s_1|D, I)}{p(s_2|D, I)} > \frac{C(u_1, s_2) - C(u_2, s_2)}{C(u_2, s_1) - C(u_1, s_1)}. \tag{297}$$

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

$$\begin{aligned}\text{posterior ratio} &= \frac{p(s_1|D, I)}{p(s_2|D, I)}\\ &= \frac{p(D_s|s_1, D_x, I)p(s_1|I)}{p(D_s|s_2, D_x, I)p(s_2|I)},\end{aligned} \tag{298}$$

where for the second equality it has been used that the normalization $p(D|I)$ cancels out between the denominator and nominator and Axiom 5 has been employed. Given there is no a priori bias towards any model, $p(s_1|I) = p(s_2|I)$

$$\text{posterior ratio} = \frac{p(D_s|s_1, D_x, I)}{p(D_s|s_2, D_x, I)}. \tag{299}$$

$p(D_s|s_1, D_x, I)$ 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 299 is referred to as Bayes factor and as a rule of thumb

Definition 61 (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 215)

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{W|D}[C(U(D), W)|D, I], \tag{300}$$

with

$$\mathbb{E}_{W|D}[C(U(D), W)|D, I] = \int dw C(U(D), w) p(w|D, I). \tag{301}$$

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 263. Picking the quadratic cost (Definition 59) yields

$$U^*(D) = \mathbb{E}_{W|D}[W|D, I] \tag{302}$$

$p(w|D, I)$ in Equation 302 can be expanded as shown in Equation 263.

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

$$p(w_B > w_A|D, I) = \int_0^1 \int_{w_A}^1 p(w_A, w_B|D, I) dw_A dw_B. \tag{303}$$

Assuming the costumer sets are independent

$$\begin{aligned} p(w_A, w_B|D, I) &= p(w_B|w_A, D, I) p(w_A|D, I) \\ &= p(w_B|D_A, I) p(w_A|D_A, I), \end{aligned} \tag{304}$$

with

$$p(w_i|D_i, I) = \frac{p(D_i|w_i, I)p(w_i|I)}{p(D_i|I)}. \tag{305}$$

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

$$p(w_i|D_i, I) = \frac{w_i^{\alpha_i - 1}(1 - w_i)^{\beta_i - 1}}{B(\alpha_i, \beta_i)}, \tag{306}$$

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 303 yields

$$p(w_B > w_A|D, I) = \sum_{j=0}^{\alpha_B - 1} \frac{B(\alpha_A + j, \beta_A + \beta_B)}{(\beta_B + j)B(1 + j, \beta_B)B(\alpha_A, \beta_A)}. \tag{307}$$

5.4 FREQUENTIST STATISTICS

Frequentist statistics is based on Definition 52, which follows the definition of objective probability (Definition 49) and the principle of fixed, unknown parameters (Axiom 3). The foundations of Frequentist statistics trace back to seminal works such as those of Neyman and Pearson [34] and Fisher [35], who laid the groundwork for much of its methodology. Subsequent developments by Wald [36], Neyman [37], and Lehmann [38] 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

$$U^*(x, w). \tag{308}$$

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

$$U^*(x, \hat{w}^*(D)), \tag{309}$$

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,

$$f : \Omega_W \times \Omega_X \to \mathbb{R}, \tag{310}$$

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 $\xi \in \Omega_\Xi$,

$$p(s|x, w, \xi, I) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(w,x)-s)^2} \tag{311}$$

where I denotes the background information (Definition 44). Here, (w, ξ) are fixed but unknown parameters. Under the quadratic cost function from Definition 59, the optimal decision rule is the conditional expectation of S given (x, w, ξ) (Theorem 13),

$$\begin{aligned} U^*(x, \hat{w}^*(D), \hat{\xi}^*(D)) &= \mathbb{E}[S|x, \hat{w}^*(D), \hat{\xi}^*(D), I] \\ &= \int s p(s|x, \hat{w}^*(D), \hat{\xi}^*(D), I) ds \\ &= f(\hat{w}^*(D), x). \end{aligned} \tag{312}$$

Equation 312 represents the Frequentist optimal decision rule conditional on the parameter estimate, whereas Equation 268 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 312 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

$$f : \Omega_W \times \Omega_X \to \Delta^K, \tag{313}$$

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 $x \in \Omega_X$

$$p(S = s|x, w, I) = f_s(w, x), \quad s \in \{1, \ldots, K\}, \tag{314}$$

with

$$\sum_{s \in \Omega_S} p(S = s|x, w, I) = 1. \tag{315}$$

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,

$$C(U(x, w), s) = 1 - \delta_{U(x,w),s} + (\lambda - 1)\delta_{U(x,w),\text{reject}}. \tag{316}$$

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

$$\begin{aligned} U^*(x, \hat{w}^*(D)) &= \arg\min_{u \in \Omega_U} \mathbb{E}[C(u, S) \mid x, \hat{w}(D), I] \\ &= \arg\min_{u \in \Omega_U} \Big(1 - f_u(\hat{w}^*(D), x) + (\lambda - 1)\delta_{u,\text{reject}}\Big). \end{aligned} \tag{317}$$

From equation Equation 317 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 62 (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 22 (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:*

$$\widehat{\mathrm{Var}}_{\mathrm{boot}}[\hat{w}] = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{w}(D^{*b}) - \overline{\hat{w}}^{*} \right)^2, \tag{318}$$

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 63 (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*

$$\begin{aligned}\mathcal{I}(w) &\equiv \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\right)^2\middle| D_x, w, I\right] \\ &= \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\middle| D_x, w, I\right].\end{aligned} \tag{319}$$

Proof. In general

$$\mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p\right)^2\right] = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p\right] + \left(\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right]\right)^2. \tag{320}$$

Now

$$\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right] = \int dD_s\left(\frac{\partial}{\partial w}\ln p\right)p \tag{321}$$

$$= \int dD_s \frac{\partial}{\partial w}p \tag{322}$$

$$= \frac{\partial}{\partial w}\int dD_s p \tag{323}$$

$$= 0, \tag{324}$$

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

$$\mathcal{I}(w) = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_s \mid D_x, w, I)\middle| D_x, w, I\right]. \tag{325}$$

□

Theorem 15 (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*

$$p(D_s|D_x, w, I) = \prod_{i=1}^{n} p(s_i|x_i, w, I). \tag{326}$$

Then, the Fisher information of the full dataset is

$$\mathcal{I}(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_s|D_x,w,I)\right)^2\middle|D_x,w,I\right] = \sum_{i=1}^{n}\mathcal{I}_i(w), \tag{327}$$

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

$$\mathcal{I}_i(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(s_i|x_i,w,I)\right)^2\middle|x_i,w,I\right]. \tag{328}$$

Definition 64 (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*

$$\hat{w}_{\text{MLE}}(D) \equiv \arg\max_{w\in\Omega_W} p(D_s|D_x,w,I). \tag{329}$$

Theorem 16 (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*

$$\sqrt{n}\left(\hat{w}_{\text{MLE}} - w\right) \xrightarrow{d} \text{Norm}\left(0, \mathcal{I}(w)^{-1}\right), \tag{330}$$

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 65 (Minimax Decision Rule)**.** *A decision rule $\hat{w}'$ is said to be minimax if it minimize the maximum expected cost, meaning (Equation 211)*

$$\begin{aligned}\hat{w}' &\equiv \inf_{\hat{w}}\sup_{w\in\Omega_W}\mathbb{E}[C(\hat{w},w)|w,I] \\ &= \inf_{\hat{w}}\sup_{w\in\Omega_W}\int dD C(\hat{w}(D),w)p(D|w,I).\end{aligned} \tag{331}$$

Theorem 17 (Mean Squared Error (MSE))**.** *The expectation of the quadratic cost function (Definition 59) can be written*

$$\begin{aligned}\mathbb{E}[C(\hat{w},w)|w,I] &= \mathbb{E}[(\hat{w}-w)^2|w,I] \\ &= \mathbb{E}[(\hat{w}-\mathbb{E}[\hat{w}|I])^2|w,I] + (w-\mathbb{E}[\hat{w}|I])^2 \\ &= \text{Var}[\hat{w}|w,I] + \text{Bias}[\hat{w}|w,I]^2\end{aligned} \tag{332}$$

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

$$\text{Bias}[\hat{w}|w, I] \equiv w - \mathbb{E}[\hat{w}|I]. \tag{333}$$

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 1 (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 59), meaning it approximately minimizes the maximum expected cost.*

Proof. From theorem Theorem 17

$$\mathbb{E}[(\hat{w} - w)^2|w, I] = \text{Var}[\hat{w}|w, I] + \text{Bias}[\hat{w}|w, I]^2. \tag{334}$$

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

$$\begin{aligned} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w, I] &\approx \text{Var}[\hat{w}_{\text{MLE}}|w, I] \\ &\approx \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \end{aligned} \tag{335}$$

where Theorem 16 was used for the second line. The Cramer-Rao lower bound [39] for variance states that

$$\text{Var}[\hat{w}|w, I] \geq \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \tag{336}$$

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

$$\sup_{w \in \Omega_W} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w, I] \approx \inf_{\hat{w}} \sup_{w \in \Omega_W} \mathbb{E}[(\hat{w} - w)^2|w, I], \tag{337}$$

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

Example 5.13.

The bias-variance decomposition (Theorem 17) 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 $S \sim \text{Ber}(w)$, 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:*

$$\hat{w}(D_s) = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{338}$$

with

$$\begin{aligned}
\mathbb{E}[\hat{w}(D_s)|w,I] &= \int dD_s \hat{w}(D_s) p(D_s|w,I) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[S|w,I] \\
&= w, \\
\text{Var}[\hat{w}(D_s)|w,I] &= \int dD_s (\hat{w}(D_s) - \mathbb{E}[\hat{w}(D_s)|w,I])^2 p(D_s|w,I) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[S|w,I] \\
&= \frac{w(1-w)}{n}, \\
\mathbb{E}[(\hat{w}(D_s) - w)^2|w,I] &= \text{Var}[\hat{w}(D_s)|w,I] + (\mathbb{E}[\hat{w}(D_s)|w,I] - w)^2 \\
&= \frac{w(1-w)}{n}.
\end{aligned} \tag{339}$$

- *Constant estimate:*

$$\hat{w} = 0.5 \tag{340}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}|w, I] &= 0.5, \\ \mathrm{Var}[\hat{w}|w, I] &= 0, \\ \mathbb{E}[(\hat{w} - w)^2|w, I] &= (0.5 - w)^2. \end{aligned} \tag{341}$$

- *First observation:*

$$\hat{w}(D_s) = s_1 \tag{342}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}(D_s)|w, I] &= \mathbb{E}[S|w, I] \\ &= w, \\ \mathrm{Var}[\hat{w}(D_s)|w, I] &= \mathrm{Var}[S|w, I] + (\mathbb{E}[\hat{w}(D_s)|w, I] - w)^2 \\ &= w(1 - w), \\ \mathbb{E}[(\hat{w}(D_s) - w)^2|w, I] &= w(1 - w). \end{aligned} \tag{343}$$

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

$$\begin{aligned} p(D_s|D_x, w, I) &= p(D_s|w, I) \\ &= \prod_{i=1}^{n} w^{s_i}(1 - w)^{1 - s_i}. \end{aligned} \tag{344}$$

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

$$\begin{aligned} \operatorname*{argmax}_w l(w) &= \operatorname*{argmax}_w p(D_s|w, I) \\ &= \operatorname*{argmax}_w \ln\left(\prod_{i=1}^{n} w^{s_i}(1 - w)^{1 - s_i}\right) \\ &= \operatorname*{argmax}_w \left[\ln w \sum_{i=1}^{n} s_i + \ln(1 - w) \sum_{i=1}^{n} (1 - s_i)\right] \end{aligned} \tag{345}$$

Now

$$\frac{d}{dw}l(w) = \frac{\sum_{i=1}^{n} s_i}{w} - \frac{n - \sum_{i=1}^{n} s_i}{1 - w} \tag{346}$$

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

$$\hat{w}_{MLE}(D_s) = \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{347}$$

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

$$\begin{aligned} p(D_s|D_x, w, I) &= p(D_s|w, I) \\ &= \prod_{i=1}^{n} we^{-ws_i}. \end{aligned} \tag{348}$$

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

$$\frac{d}{dw}l(w) = \frac{n}{w} - \sum_{i=1}^{n} s_i \tag{349}$$

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

$$\hat{w}_{MLE}(D_s) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} s_i}. \tag{350}$$

APPENDIX A

Hamiltonian Monte Carlo

This appendix is taken from Petersen [40]. The Hamiltonian Monte Carlo Algorithm (HMC algorithm) is a Markov Chain Monte Carlo (MCMC) algorithm used to evaluate integrals on the form

$$\begin{aligned}\mathbb{E}[f] &= \int f(\theta)g(\theta)d\theta \\ &\approx \frac{1}{N}\sum_{j\in g} f(\theta_j),\end{aligned} \tag{351}$$

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 [28]

$$g(\theta') = \int p(\theta'|\theta)g(\theta)d\theta. \tag{352}$$

Equation 352 is implied by the stronger condition of detailed balance, defined viz

$$p(\theta'|\theta)g(\theta) = p(\theta|\theta')g(\theta'). \tag{353}$$

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 [41, 42]. 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 [27] 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 [29] 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

$$g = e^{\frac{F-H}{k_B T}} \Rightarrow H = F - k_B T \ln[g], \tag{354}$$

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

$$H = -k_B T \left[\ln[g] - \sum_i \frac{p_i^2}{2m_i} \right] + const, \tag{355}$$

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 [29], however, the temperature can be used to manipulate the range of states which can be accepted e.g. via simulated annealing [44]. Here $T = k_b^{-1}$ will be adopted in accordance with [28, 29] and as such

$$H = \sum_i \frac{p_i^2}{2m_i} - \ln[g]. \tag{356}$$

The dynamics in parameter space are determined by Hamiltons equations

$$\dot{\theta}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial \theta_i}, \tag{357}$$

with θ_i denoting the different variables (coefficients). In order to implement Hamiltons equations, they are discretized via the leap frog method [28, 29] viz

$$\begin{aligned} p_i\left(t + \frac{\epsilon}{2}\right) &= p_i(t) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t), p_i(t))}{\partial \theta_i}, \\ \theta_i(t+\epsilon) &= \theta_i(t) + \frac{\epsilon}{m_i} p_i\left(t + \frac{\epsilon}{2}\right), \\ p_i(t+\epsilon) &= p_i\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t+\frac{\epsilon}{2}), p_i(t+\frac{\epsilon}{2}))}{\partial \theta_i}, \end{aligned} \tag{358}$$

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

$$\mathbb{P}(\text{transition}) = \min\left[1, e^{H_{initial} - H_{final}}\right]. \tag{359}$$

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 $\mathcal{O}(\epsilon^2)$ [28] 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

$$H = \frac{q^2}{2\sigma^2} + \frac{p^2}{2}. \tag{360}$$

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

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \begin{bmatrix} 1 - \frac{\epsilon^2}{2\sigma^2} & \epsilon \\ \epsilon(\frac{1}{4}\epsilon^2\sigma^{-4} - \sigma^{-2}) & 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \tag{361}$$

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

$$\text{Eigenvalues} = 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \pm \epsilon\sigma^{-1}\sqrt{\frac{1}{4}\epsilon^2\sigma^{-2} - 1}. \tag{362}$$

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

$$\epsilon < 2\sigma. \tag{363}$$

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 [43] a stable solution with $\frac{1}{2}p^T\Sigma^{-1}p$ as the kinetic term in the Hamiltonian require

$$\epsilon_i < 2\lambda_i^{-\frac{1}{2}}, \tag{364}$$

for each eigenvalue λ_i of the matrix

$$M_{ij} = (\Sigma^{-1})_{ij}\frac{\partial^2 H}{\partial q_i \partial q_j}, \tag{365}$$

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

$$\epsilon_i < 2\sqrt{\frac{m_i}{\frac{\partial^2 H}{\partial q_i^2}}}. \tag{366}$$

Setting ϵ according to Equation 364 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 [28, 29] (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] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: http://probml.github.io/book2.

[2] Steven M. Lavalle. *Planning Algorithms*. Cambridge University Press, 2006. ISBN: 0521862051.

[3] D. S. Sivia and J. Skilling. *Data Analysis - A Bayesian Tutorial*. 2nd. Oxford Science Publications. Oxford University Press, 2006.

[4] S.H. Chan. *Introduction to Probability for Data Science*. Michigan Publishing, 2021. ISBN: 9781607857464. URL: https://books.google.dk/books?id=GR2jzgEACAAJ.

[5] Alexander Drewitz. *Introduction to Probability and Statistics*. Preliminary version, February 1. University of Cologne, 2019.

[6] J. NavrÃątil. "Radon-Nikodym theorem in spaces of measures." In: *Mathematica Scandinavica* 48.1981 (1981), pp. 5–12. URL: https://www.mscand.dk/article/view/11916.

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

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

[9] E. T. Jaynes. "Marginalization and Prior Probabilities." In: *Bayesian Analysis in Econometrics and Statistics*. Ed. by A. Zellner. Amsterdam: North-Holland Publishing Company, 1980.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] Edward E. Leamer. *Specification Searches: Ad Hoc Inference with Non-experimental Data*. Wiley, 1978, p. 25.

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

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

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

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

[30] 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/.

[31] 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.

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

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

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

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

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

[41] 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.

[42] 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.

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

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

Index