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 2.1 (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.*

$$A = \{a_1, a_2, a_3\}. \tag{2.1}$$

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

$$A = \{x \mid x \text{ is a natural number}\}. \tag{2.2}$$

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

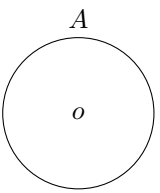

Figure 1: The graphical representation of a generic set A with generic element o.

Definition 2.2 (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 2.3 (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

$$\begin{aligned} A \times B = \{&(a_1, b_1), (a_1, b_2), (a_1, b_3), \\ &(a_2, b_1), (a_2, b_2), (a_2, b_3)\}. \end{aligned} \tag{2.3}$$

Definition 2.4 (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,*

$$A \subseteq B \quad \iff \quad \forall x \in A \colon x \in B. \tag{2.4}$$

By this definition, a set is always a subset of itself.

Definition 2.5 (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.*

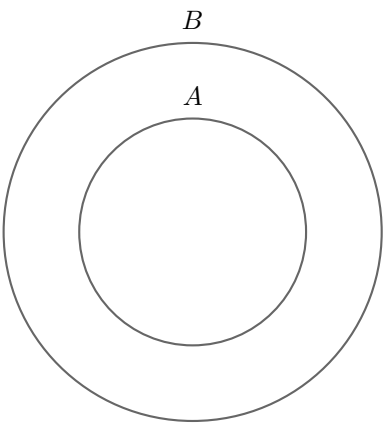

Figure 2: The graphical representation of $A \subset B$.

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

Definition 2.7 (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*

$$2^A = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, A\}. \tag{2.5}$$

Definition 2.8 (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 2.9 (Closure)**.** *The set* A *is said to be closed under a certain operation if, for every pair of elements* $x, y \in A$*, the result of applying the operation to* x y *is also in* A*.*

Definition 2.10 (Closure)**.** *A set* A *is said to be* closed under an operation $*$ *if, for all* $x, y \in A$*, the result of applying* $*$ *to* x *and* y *is also an element of* A*; that is,*

$$x, y \in A \;\Rightarrow\; x * y \in A. \tag{2.6}$$

Example 2.5.
Let $\mathcal{E} = \{2k \mid k \in \mathbb{Z}\}$ *denote the set of even integers. Then* $\mathcal{E}$ *is closed under addition, since for any* $x, y \in \mathcal{E}$*,*

$$x + y \in \mathcal{E}. \tag{2.7}$$

In contrast, the set of odd integers $\mathbb{O} = \{2k + 1 \mid k \in \mathbb{Z}\}$ *is not closed under addition, because* $x + y$ *is even when* $x, y \in \mathbb{O}$*.*

Definition 2.11 (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 $A \cup B$.*

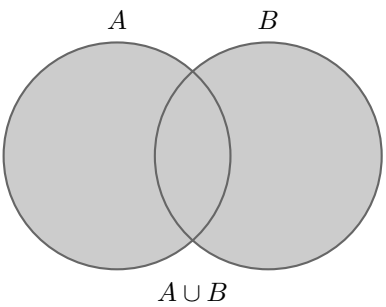

Figure 3: The graphical representation of $A \cup B$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 2.12 (Finite and Infinite Unions)**.** *For a collection of sets $\{A_i\}$, 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 2.13 (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 $A \cap B$.*

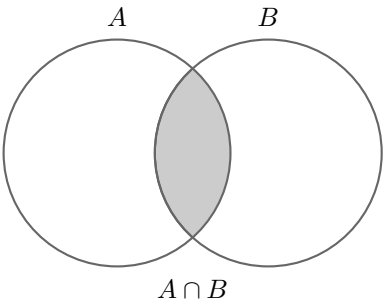

Figure 4: The graphical representation of $A \cap B$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 2.14 (Finite and Infinite Intersections)**.** *For a collection of sets* $\{A_i\}$*, 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 2.15 (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* $A \cap B = \emptyset$*.*

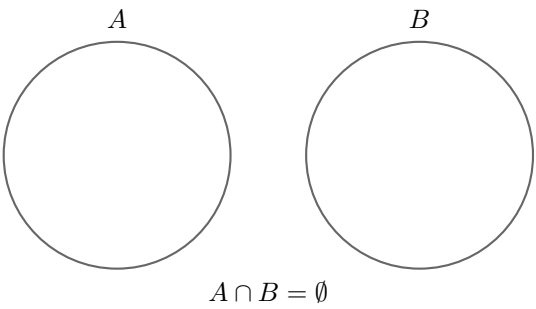

Figure 5: The graphical representation of $A \cap B = \emptyset$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 2.16 (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* $(A \cap B)^c$*.*

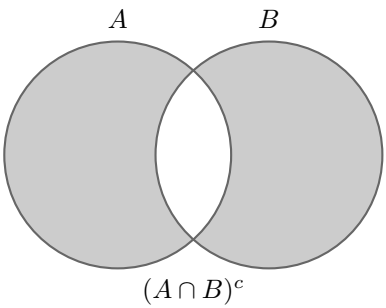

Figure 6: The graphical representation of $(A \cap B)^c$. Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 2.17 (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 $B \setminus A$.*

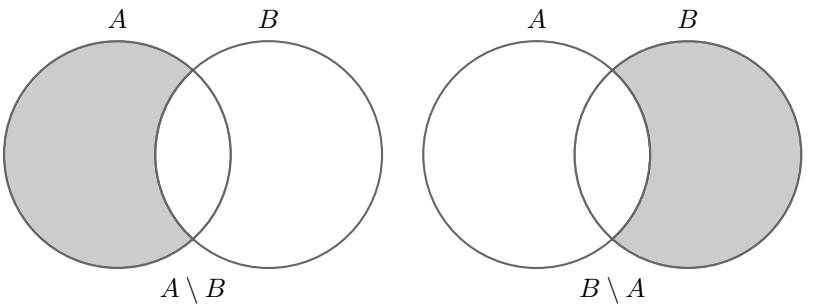

Figure 7: The graphical representation of $A \setminus B$ (left) and $B \setminus A$ (right). Each circle represents the sets and the colored region represents the result of the binary operation.

Definition 2.18 (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 2.19 (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.

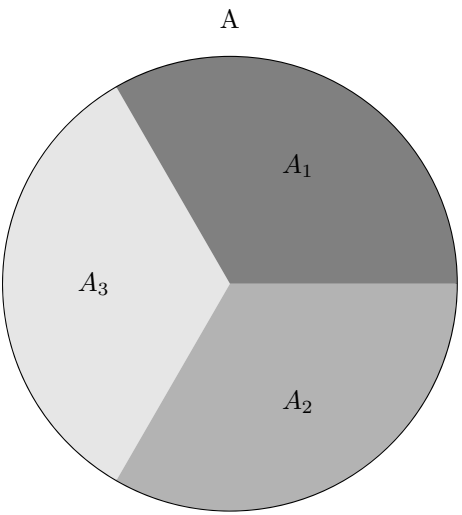

Figure 8: A graphical representation of $A = \{A_1, A_2, A_3\}$, where the A_j are the disjoint subsets forming a partition of A.

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 3.1 (Sample Space)**.** *The sample space, denoted by Ω, is the set of all possible outcomes of a random experiment. It serves as the universal set (see Definition 2.8) for the experiment, providing the foundation for defining probabilities of events.*

Definition 3.2 (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.*

Definition 3.3 (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 3.1.

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.

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

$$\mathcal{G}_{trivial} = \{\emptyset, \Omega\}. \tag{3.1}$$

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 3.4 (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 3.5 (Event Space)**.** *The event space, denoted by $\mathcal{F}$, is the collection of all subsets of the sample space Ω that are considered valid events for a random experiment. Formally, $\mathcal{F}$ is required to be a σ-algebra, ensuring it is closed under complementation, countable unions, and countable intersections.*

Remark 3.1 (Typical Event Spaces)**.** *For a discrete sample space Ω, $\mathcal{F}$ is typically the power set of Ω (see Definition 2.7). For a continuous sample space, $\mathcal{F}$ is typically the Borel σ-algebra (see Definition 3.4), 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

$$\begin{aligned}\mathcal{F} &= \{\emptyset, \{\boxdot\}, \{\boxdot, \boxdot\}, \{\boxdot\}, \{\boxdot, \boxdot, \boxdot, \boxdot\}, \{\boxdot\}, \ldots, \Omega\} \\ &= 2^{\Omega}.\end{aligned} \tag{3.2}$$

Definition 3.6 (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 corresponding event space.*

Definition 3.7 (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*

$$\mu \colon \mathcal{F} \to [0, \infty] \tag{3.3}$$

that satisfies Axiom 3.1 (non-negativity) and Axiom 3.2 (additivity).

Axiom 3.1 (Non-negativity)**.** *For any event $E \in \mathcal{F}$, the measure $\mu(E)$ is non-negative, satisfying*

$$\mu(E) \geq 0 \quad \forall E \in \mathcal{F}. \tag{3.4}$$

Axiom 3.2 (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*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i) \quad \forall E_i \in \mathcal{F} \text{ where } \bigcap_{i=1}^{\infty} E_i = \emptyset. \tag{3.5}$$

Definition 3.8 (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*

$$\Omega = \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \mu(A_i) < \infty \ \ \forall i \in \mathbb{N}. \tag{3.6}$$

Definition 3.9 (Measurable Function)**.** *Let $(\Omega, \mathcal{F})$ and $(\Omega_X, \mathcal{F}_X)$ be measurable spaces. A function*

$$X \colon \Omega \to \Omega_X \tag{3.7}$$

is said to be measurable if

$$X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{F}_X, \tag{3.8}$$

where

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}. \tag{3.9}$$

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

Definition 3.10 (Lebesgue Measure)**.** *Let $(\Omega, \mathcal{F})$ be a continuous measurable space, where Ω is the sample space and $\mathcal{F} = \mathcal{B}(\mathbb{R})$ is the event space. The Lebesgue measure λ is a measure, in the sense of Definition 3.7, defined on $\mathcal{F}$ such that, for every interval $(a, b] \subseteq \mathbb{R}$,*

$$\lambda((a, b]) = b - a. \tag{3.10}$$

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

Definition 3.11 (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 3.7, defined on $\mathcal{F}$ such that for every event $E \in \mathcal{F}$,*

$$\nu(E) = |E|, \tag{3.11}$$

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 3.12 (Probability Measure)**.** *Let* $(\Omega, \mathcal{F})$ *be a measurable space, where* Ω *is the sample space and* $\mathcal{F}$ *is the event space. A probability measure* $\mathbb{P}$ *is a measure, in accordance with Definition 3.7, defined on* $(\Omega, \mathcal{F})$ *that, in addition to satisfying Axiom 3.1 (non-negativity) and Axiom 3.2 (additivity), satisfies the normalization property*

$$\mathbb{P}(\Omega) = 1. \tag{3.12}$$

Probability can be interpreted as assigning a number between 0 *and* 1 *to each event in a mathematically consistent way [4].*

Definition 3.13 (Objective Probability Measure)**.** *Let* $\mathbb{P}$ *denote a probability measure defined on the measurable space* $(\Omega, \mathcal{F})$*. Interpreting the probability measure as an objective probability measure [5, 6] consists in viewing* $\mathbb{P}(E)$ *as the long-run relative frequency of the event* $E \in \mathcal{F}$ *over repeated independent trials, provided that the limit exists.*

Definition 3.14 (Subjective Probability Measure)**.** *Let* $\mathbb{P}$ *denote a probability measure defined on the measurable space* $(\Omega, \mathcal{F})$*. Interpreting the probability measure as a subjective probability measure [7, 8] consists in assigning values to events based on rational degrees of belief derived from prior knowledge, constraints, or partial information, in a way that avoids internal inconsistencies (i.e., satisfies Dutch-book coherence [9]).*

Definition 3.15 (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 3.2 (Events for Continuous Sample Spaces)**.** *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*

$$E = (0.2, 0.5) \tag{3.13}$$

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

Remark 3.3 (Reason for Borel σ-algebra)**.** *The restriction to Borel sets in continuous sample spaces avoids paradoxical constructions such as non-measurable sets (e.g., the Vitali set [10]), which cannot be consistently assigned a probability [11], ensuring that the probability measure is well defined on all events in* $\mathcal{F}$ *[12].*

Example 3.4.
Consider choosing a real number uniformly at random from the interval $[0,1]$*. Here the sample space is* $\Omega = [0,1]$*. The event space* $\mathcal{F}$ *cannot 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 3.16 (Independence)**.** *Two events* $E_1, E_2 \in \mathcal{F}$ *in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *are said to be independent if and only if*

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2). \tag{3.14}$$

Definition 3.17 (Conditional Probability)**.** *Let* $E_1, E_2 \in \mathcal{F}$ *be events 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 by*

$$\mathbb{P}(E_1 \mid E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}. \tag{3.15}$$

Definition 3.18 (Conditional Independence)**.** *Events* E_1 *and* E_2 *are conditionally independent given* $E_3 \in \mathcal{F}$ *if*

$$\mathbb{P}(E_1 \cap E_2 \mid E_3) = \mathbb{P}(E_1 \mid E_3)\,\mathbb{P}(E_2 \mid E_3), \tag{3.16}$$

provided $\mathbb{P}(E_3) > 0$*.*

Theorem 3.1 (Chain Rule)**.** *Let* $E_1, E_2, E_3 \in \mathcal{F}$ *be events in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *with* $\mathbb{P}(E_3) > 0$*. Then the chain rule states that*

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_1 \mid E_2 \cap E_3)\,\mathbb{P}(E_2 \mid E_3)\,\mathbb{P}(E_3). \tag{3.17}$$

Proof. From the definition of conditional probability in Definition 3.17

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_1 | E_2 \cap E_n)\mathbb{P}(E_2 \cap E_3). \tag{3.18}$$

Using the definition of conditional probability again

$$\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(E_2|E_3)\mathbb{P}(E_3). \tag{3.19}$$

which leads to Theorem 3.1. □

Theorem 3.2 (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*

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)}. \tag{3.20}$$

Proof. Bayes theorem follows directly from applying Theorem 3.1 and applying the concept of symmetry as follows

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &= \mathbb{P}(E_1|E_2)\mathbb{P}(E_2) \\ &= \mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\end{aligned} \tag{3.21}$$

from which

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)} \tag{3.22}$$

which is Theorem 3.2. □

Theorem 3.3 (Law of Total Probability / Marginalization)**.** *Let* $\{E_1, E_2, \ldots, E_n\}$ *be a finite partition, in the sense of Definition 2.19, of the sample space* Ω *in a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. Then, for any event* $A \in \mathcal{F}$*,*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap E_i). \tag{3.23}$$

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

$$\sum_i \mathbb{P}(A \cap E_i) = \mathbb{P}\Big(\bigcup_i (A \cap E_i)\Big). \tag{3.24}$$

$\bigcup_i (A \cap E_i)$ 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

$$\begin{aligned}\bigcup_i (A \cap E_i) &= (A, \bigcup_i E_i) \\ &= (A \cap \Omega) \\ &= A.\end{aligned} \tag{3.25}$$

Combining Equation 3.24 and Equation 3.25 then yields

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap E_i) \tag{3.26}$$

which is Theorem 3.3. □

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 3.17*

$$\begin{aligned} \mathbb{P}(E_1|E_2) &= \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \\ &= 1 \end{aligned} \tag{3.27}$$

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 3.19 (Random Variable)**.** *A random variable* X *is a measurable function in the sense of Definition 3.9,*

$$X \colon \Omega \to \Omega_X \tag{3.28}$$

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 3.4 (Types of Random Variables)**.** *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 3.20 (Image Measure)**.** *Let*

$$X \colon \Omega \to \Omega_X \tag{3.29}$$

be a random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *to a measurable space* $(\Omega_X, \mathcal{F}_X)$. *Then [13]*

$$\mathbb{P} \circ X^{-1} \colon \mathcal{F}_X \to [0, 1] \tag{3.30}$$

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 3.5 (Maginalization via Random Variable)**.** *Theorem 3.3 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*

$$X \colon \Omega \to \Omega_X \tag{3.31}$$

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 3.3 can be written rigorously in terms of the image measure* $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ *as*

$$\mathbb{P}(A) = \int_{\Omega_X} \mathbb{P}(A \mid X^{-1}(\{x\}))\, d\mathbb{P}_X(x), \tag{3.32}$$

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

Definition 3.21 (Expected value)**.** *Let*

$$X \colon \Omega \to \Omega_X \tag{3.33}$$

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

$$\mathbb{P}_X = \mathbb{P} \circ X^{-1} \tag{3.34}$$

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*

$$\mathbb{E}_X[X] \equiv \int_{\Omega_X} x \mathrm{d}\mathbb{P}_X(x). \tag{3.35}$$

Theorem 3.4 (Non-negativity of expected value)**.** *Let*

$$X \colon \Omega \to \Omega_X \tag{3.36}$$

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 3.21

$$\mathbb{E}_X[X] = \int_{\Omega_X} x \mathrm{d}\mathbb{P}_X(x), \tag{3.37}$$

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 3.5 (Linearity of expected value)**.** *Let*

$$X \colon \Omega \to \Omega_X \tag{3.38}$$

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 3.21

$$\begin{aligned}\mathbb{E}_X[a+X] &= \int_{\Omega_X} (a+x)\mathrm{d}\mathbb{P}_X(x) \\ &= a\int_{\Omega_X} \mathrm{d}\mathbb{P}_X(x) + \int_{\Omega_X} x\mathrm{d}\mathbb{P}_X(x) \\ &= a + \mathbb{E}_X[X],\end{aligned} \tag{3.39}$$

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

$$\begin{aligned}\mathbb{E}_X[aX] &= \int_{\Omega_X} ax\mathrm{d}\mathbb{P}_X(x) \\ &= a\int_{\Omega_X} x\mathrm{d}\mathbb{P}_X(x) \\ = a\mathbb{E}_X[X].\end{aligned} \tag{3.40}$$

□

Remark 3.6 (Law of the Unconscious Statistician)**.** *Let*

$$X\colon \Omega \to \Omega_X \tag{3.41}$$

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

$$g\colon \Omega_X \to \mathbb{R} \tag{3.42}$$

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

$$\mathbb{E}_X[g(X)] \equiv \int_{\Omega_X} g(x)\mathrm{d}\mathbb{P}_X(x). \tag{3.43}$$

Definition 3.22 (Variance)**.** *Let*

$$X\colon \Omega \to \Omega_X \tag{3.44}$$

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*

$$\begin{aligned}\mathrm{Var}_X[X] &\equiv \mathbb{E}_X[(X - \mathbb{E}_X[X])^2] \\ &= \mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2.\end{aligned} \tag{3.45}$$

Theorem 3.6 (Markov's Inequality)**.** *Let*

$$X\colon \Omega \to \Omega_X \tag{3.46}$$

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*

$$\mathbb{P}_X([a,\infty)) \leq \frac{\mathbb{E}_X[X]}{a}. \tag{3.47}$$

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

$$a\, 1_{[a,\infty)}(X(\omega)) \leq X(\omega), \quad \forall \omega \in \Omega. \tag{3.48}$$

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

$$a\, \mathbb{E}_X[1_{[a,\infty)}] \leq \mathbb{E}_X[X]. \tag{3.49}$$

By definition of the image measure,

$$\begin{aligned} \mathbb{E}_X[1_{[a,\infty)}] &= \int 1_{[a,\infty)}(x) \mathrm{d}\mathbb{P}_X(x) \\ &= \mathbb{P}_X([a,\infty)). \end{aligned} \tag{3.50}$$

Hence,

$$a\mathbb{P}_X([a,\infty)) \leq \mathbb{E}_X[X], \tag{3.51}$$

and dividing both sides by a yields the inequality. □

Definition 3.23 (Probability Density with Respect to a Measure)**.** *Let*

$$X\colon \Omega \to \Omega_X \tag{3.52}$$

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 3.8. 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 [14] there exists a measurable function*

$$p_X = \frac{d\mathbb{P}_X}{d\mu}, \tag{3.53}$$

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

$$\mathbb{P}_X(B) = \int_B p_X(x) \mathrm{d}\mu(x), \quad \forall B \in \mathcal{F}_X. \tag{3.54}$$

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

$$p_X(x) \geq 0, \quad \textit{and} \quad \int_{\Omega_X} p_X(x) \mathrm{d}\mu(x) = 1. \tag{3.55}$$

Remark 3.7 (Probability Density Function)**.** *If $\mu = \lambda$ is the Lebesgue measure (Definition 3.10) on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the probability density p_X is called the probability density function (PDF). For $\mu = \lambda$, Definition 3.23 gives*

$$\mathbb{P}_X(B) = \int_B p_X(x) \mathrm{d}\lambda(x), \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{3.56}$$

Remark 3.8 (Probability Mass Function)**.** *If $\mu = \nu$ is the counting measure (Definition 3.11) on Ω_X, then the probability density p_X is called the probability mass function (PMF). For $\mu = \nu$, Definition 3.23 gives*

$$\begin{aligned} \mathbb{P}_X(B) &= \int_B p_X(x) \mathrm{d}\nu(x) \\ &= \sum_{x \in B} p_X(x). \end{aligned} \tag{3.57}$$

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

$$\mathbb{P}_X(\{x\}) = p_X(x). \tag{3.58}$$

Remark 3.9 (Expected value of a discrete random variable)**.** *If X is a discrete random variable with probability mass function p_X, then the expected value reduces to*

$$\mathbb{E}_X[X] = \sum_{x \in \Omega_X} x p_X(x). \tag{3.59}$$

Equation 3.59 follows directly from Definition 3.21, since the image measure $\mathbb{P}_X$ is concentrated on singletons $\{x\}$ in the discrete case.

Remark 3.10 (Expected value of a continuous random variable)**.** *Let X be a continuous random variable with PDF p_X on $\Omega_X \subseteq \mathbb{R}$. From Definition 3.21 and Definition 3.23*

$$\mathbb{E}_X[X] = \int_{\Omega_X} x p_X(x) \mathrm{d}\lambda(x), \tag{3.60}$$

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

$$\mathbb{E}_X[X] = \int_{\Omega_X} x p_X(x) \mathrm{d}x. \tag{3.61}$$

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

Example 3.6.
Let X be a continuous random variable with PDF p_X on $\Omega_X \subseteq \mathbb{R}$. For the interval (event) $[a,b] \subseteq \Omega_X$,

$$\begin{aligned}\mathbb{P}(X^{-1}([a,b])) &= \mathbb{P}_X([a,b]) \\ &= \int_{[a,b]} p_X(x)\mathrm{d}\lambda(x) \\ &= \int_a^b p_X(x)\mathrm{d}x.\end{aligned} \tag{3.62}$$

Definition 3.24 (Joint Probability Measure)**.** *Let*

$$X\colon \Omega \to \Omega_X, \quad Y\colon \Omega \to \Omega_Y \tag{3.63}$$

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

$$\mathbb{P}_{X,Y} = \mathbb{P} \circ (X,Y)^{-1} \tag{3.64}$$

defined on the measurable space

$$(\Omega_{X_1} \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y). \tag{3.65}$$

All probability measures related to the random variables can be derived from the joint probability measure via Theorem 3.3.

Remark 3.11 (Marginalization from a Joint Measure)**.** *Let*

$$X\colon \Omega \to \Omega_X, \quad Y\colon \Omega \to \Omega_Y \tag{3.66}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to measurable spaces $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$. 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 3.3

$$\mathbb{P}_X(A) = \mathbb{P}_{X,Y}(A \times \Omega_Y). \tag{3.67}$$

Theorem 3.7 (Law of total expectation)**.** *Let*

$$X\colon \Omega \to \Omega_X, \quad Y\colon \Omega \to \Omega_Y \tag{3.68}$$

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

$$\mathbb{E}_X[X] = \mathbb{E}_Y\big[\mathbb{E}_{X|Y}[X \mid Y]\big]. \tag{3.69}$$

Proof. Let

$$X\colon \Omega \to \Omega_X, \quad Y\colon \Omega \to \Omega_Y \tag{3.70}$$

be random variables from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the 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 3.17,

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} x \mathrm{d}\mathbb{P}_X(x) \\
&= \int_{\Omega_X} \int_{\Omega_Y} x \mathrm{d}\mathbb{P}_{X,Y}(x,y) \\
&= \int_{\Omega_Y} \int_{\Omega_X} x \mathrm{d}\mathbb{P}_{X|Y}(x) \mathrm{d}\mathbb{P}_Y(y) \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X \mid Y]].
\end{aligned} \tag{3.71}$$

or equivalently in terms of the probability densities

$$\begin{aligned}
\mathbb{E}_X[X] &= \int_{\Omega_X} x p_X(x) \mathrm{d}\mu_X(x) \\
&= \int_{\Omega_Y} \int_{\Omega_X} x p_{X,Y}(x,y) \mathrm{d}\mu_X(x) \mathrm{d}\mu_Y(y) \\
&= \int_{\Omega_Y} p_Y(y) \left(\int_{\Omega_X} x p_{X|Y}(x \mid y) \mathrm{d}\mu_X(x) \right) \mathrm{d}\mu_Y(y) \\
&= \mathbb{E}_Y[\mathbb{E}_{X|Y}[X \mid Y]].
\end{aligned} \tag{3.72}$$

□

Theorem 3.8 (Expectation of product of independent random variables)**.** *Let $X\colon \Omega \to \Omega_X$ and $Y\colon \Omega \to \Omega_Y$ 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*

$$\mathbb{E}_{X,Y}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]. \tag{3.73}$$

Proof. If the random variables are independent, then according to Definition 3.16

$$\mathbb{P}_{X,Y}(\{x,y\}) = \mathbb{P}_X(\{x\})\mathbb{P}_Y(\{y\}) \tag{3.74}$$

meaning

$$\begin{aligned}\mathbb{E}_{X,Y}[XY] &= \int_{\Omega_X}\int_{\Omega_Y} xy\mathrm{d}\mathbb{P}_{X,Y}(x,y)\\ &= \int_{\Omega_X} x\mathrm{d}\mathbb{P}_X(x)\int_{\Omega_Y} y\mathrm{d}\mathbb{P}_Y(y)\\ &= \mathbb{E}_X[X]\mathbb{E}_Y[Y].\end{aligned} \tag{3.75}$$

□

Definition 3.25 (Covariance)**.** *Let $X\colon \Omega \to \Omega_X$ and $Y\colon \Omega \to \Omega_Y$ 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*

$$\begin{aligned}\mathrm{Cov}_{X,Y}[X,Y] &= \mathbb{E}_{X,Y}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])]\\ &= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y],\end{aligned} \tag{3.76}$$

Theorem 3.9 (Covariance of independent random variables)**.** *Let $X\colon \Omega \to \Omega_X$ and $Y\colon \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If X and Y are independent, then their covariance is*

$$\mathrm{Cov}_{X,Y}[X,Y] = 0. \tag{3.77}$$

Proof. Using Theorem 3.8 in Definition 3.25 yields $\mathrm{Cov}_{X,Y}[X,Y] = 0$. □

Definition 3.26 (Correlation)**.** *Let $X\colon \Omega \to \Omega_X$ and $Y\colon \Omega \to \Omega_Y$ 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*

$$\begin{aligned}\mathrm{Corr}_{X,Y}[X,Y] &= \frac{\mathrm{Cov}_{X,Y}[X,Y]}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_Y[Y]}}\\ &= \frac{\mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]}{\sqrt{(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2)\,(\mathbb{E}_Y[Y^2] - \mathbb{E}_Y[Y]^2)}}.\end{aligned} \tag{3.78}$$

Remark 3.12 (Correlation vs. Covariance)**.** *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 3.27 (Change of Variables for PDFs)**.** *Let $X\colon \Omega \to \Omega_X$ be a continuous random variable with probability density function (PDF) p_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 p_Y, is given by the change of variables formula [3]*

$$p_Y(y) = p_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right|, \quad y \in \Omega_Y, \tag{3.79}$$

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

Example 3.7.

Let $X\colon \Omega \to \Omega_X$ and $Y\colon \Omega \to \Omega_Y$ be continuous random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Using Definition 3.22 and Definition 3.25, the variance of a sum can be written

$$\begin{aligned} \operatorname{Var}_{X,Y}[X+Y] &= \mathbb{E}_{X,Y}[(X+Y-\mathbb{E}_{X,Y}[X+Y])^2] \\ &= \mathbb{E}_X[(X-\mathbb{E}_X[X])^2] + \mathbb{E}_Y[(Y-\mathbb{E}_Y[Y])^2] \\ &\quad + 2\mathbb{E}_{X,Y}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])] \\ &= \operatorname{Var}_X[X] + \operatorname{Var}_Y[Y] + 2\operatorname{Cov}_{X,Y}[X,Y]. \end{aligned} \tag{3.80}$$

Example 3.8.

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

$$\begin{aligned} Y &= g(X) \\ &= aX + b, \end{aligned} \tag{3.81}$$

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

$$g^{-1}(y) = \frac{y-b}{a}. \tag{3.82}$$

Using Definition 3.27, the PDF of Y is

$$\begin{aligned} p_Y(y) &= p_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= p_X\left(\frac{y-b}{a}\right) \left| \frac{1}{a} \right|. \end{aligned} \tag{3.83}$$

Hence,

$$p_Y(y) = \frac{1}{|a|} p_X\left(\frac{y-b}{a}\right). \tag{3.84}$$

Example 3.9.

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

$$\begin{aligned} Y &= g(X) \\ &= \frac{e^X}{1+e^X}. \end{aligned} \tag{3.85}$$

The inverse function is

$$g^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{3.86}$$

Using Definition 3.27, the PDF of Y is

$$\begin{aligned} p_Y(y) &= p_X\big(g^{-1}(y)\big) \left|\frac{d}{dy} g^{-1}(y)\right| \\ &= \text{const} \cdot \left|\frac{d}{dy} \ln\left(\frac{y}{1-y}\right)\right| \\ &= \text{const} \cdot \frac{1}{y(1-y)}, \quad y \in (0,1). \end{aligned} \tag{3.87}$$

Theorem 3.10 (Error Propagation)**.** *Let*

$$X_1\colon \Omega \to \Omega_{X_1}, \ldots, X_n\colon \Omega \to \Omega_{X_n} \tag{3.88}$$

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

$$g\colon \Omega_{X_1} \times \cdots \times \Omega_{X_n} \to \mathbb{R} \tag{3.89}$$

be a differentiable function of these variables. Denote for shorthand

$$X = (X_1, \ldots, X_n) \tag{3.90}$$

and

$$\mathbb{E}_X[X] = (\mathbb{E}_{X_1}[X_1], \ldots, \mathbb{E}_{X_n}[X_n]). \tag{3.91}$$

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

$$\begin{aligned}\mathrm{Var}_X[g(X)] = &\sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \\ &+ \sum_{i \neq j} \frac{\partial g(X)}{\partial X_i} \frac{\partial g(X)}{\partial X_j}\bigg|_{X=\mathbb{E}_X[X]} \mathrm{Cov}_{X_i,X_j}[X_i, X_j] \\ &+ \mathcal{O}(\|X - \mathbb{E}_X[X]\|^3).\end{aligned} \tag{3.92}$$

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

$$\begin{aligned}g(X) =& g(\mathbb{E}_X[X]) + \sum_{i=1}^{n} \frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} (X_i - \mathbb{E}_{X_i}[X_i]) \\ &+ \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2).\end{aligned} \tag{3.93}$$

Consequently

$$\begin{aligned}\mathbb{E}_X[g(X)] &= g(\mathbb{E}_X[X]) + \sum_{i=1}^{n} \frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \cancelto{0}{\mathbb{E}_{X_i}[X_i - \mathbb{E}_{X_i}[X_i]]} \\ &\quad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2) \\ &= g(\mathbb{E}_X[X]) + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2)\end{aligned} \tag{3.94}$$

meaning the variance of g can be approximated as follows

$$\begin{aligned}\mathrm{Var}_X[g(X)] &= \mathbb{E}_X\left[(g(X) - \mathbb{E}_X[g(X)])^2\right] \\ &= \mathbb{E}_X\Bigg[\Bigg(\sum_{i=1}^{n} (X_i - \mathbb{E}_{X_i}[X_i]) \frac{\partial g}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \\ &\qquad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^2)\Bigg)^2\Bigg] \\ &= \sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\bigg|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \\ &\quad + \sum_{i \neq j} \frac{\partial g(X)}{\partial X_i} \frac{\partial g(X)}{\partial X_j}\bigg|_{X=\mathbb{E}_X[X]} \mathrm{Cov}_{X_i,X_j}[X_i, X_j] \\ &\quad + \mathcal{O}(\|X - \mathbb{E}_X[X]\|^3).\end{aligned} \tag{3.95}$$

□

Remark 3.13 (Error propagation for independent Random Variables)**.** *Let*

$$X_1 \colon \Omega \to \Omega_{X_1}, \ldots, X_n \colon \Omega \to \Omega_{X_n} \tag{3.96}$$

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 3.9* $\mathrm{Cov}_{X_i,X_j}[X_i, X_j] = 0 \ \forall i \neq j$*. In this case, Theorem 3.10 simplifies to*

$$\mathrm{Var}_X[g(X)] \approx \sum_{i=1}^{n} \left(\frac{\partial g(X)}{\partial X_i}\Big|_{X=\mathbb{E}_X[X]} \right)^2 \mathrm{Var}_{X_i}[X_i] \tag{3.97}$$

where

$$X = (X_1, \ldots, X_n) \tag{3.98}$$

and

$$\mathbb{E}_X[X] = (\mathbb{E}_{X_1}[X_1], \ldots, \mathbb{E}_{X_n}[X_n]). \tag{3.99}$$

Equation 3.97 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

$$X \sim \mathrm{Norm}(2m, (0.01m)^2), \quad Y \sim \mathrm{Norm}(3m, (0.02m)^2). \tag{3.100}$$

The variance of the area XY *can be determined exactly from Theorem 3.10*

$$\begin{aligned} \mathrm{Var}_{X,Y}[XY] &= \mathbb{E}_{X,Y}[(XY)^2] - (\mathbb{E}_{X,Y}[XY])^2 \\ &= (\mathrm{Var}_X[X] + \mathbb{E}_X[X]) \Big(\mathrm{Var}_Y[Y] + \mathbb{E}_Y[Y]) - \mathbb{E}_X[X]^2 \mathbb{E}_Y[Y]^2 \\ &= \mathbb{E}_Y[Y]^2 \, \mathrm{Var}_X[X] + \mathbb{E}_X[X]^2 \, \mathrm{Var}_Y[Y] + \mathrm{Var}_X[X] \, \mathrm{Var}_Y[Y] \end{aligned} \tag{3.101}$$

where Theorem 3.8 has been applied. Via the linear approximation from Remark 3.13 the variance of the area can be approximated as follows

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

(3.102)

Comparing Equation 3.101 and Equation 3.102 the relative difference can be written

$$\frac{\mathrm{Var}_{X,Y}[XY]|_{\textit{linear approx.}} - \mathrm{Var}_{X,Y}[XY]}{\mathrm{Var}_{X,Y}[XY]} = -\frac{\mathrm{Var}_X[X]\,\mathrm{Var}_Y[Y]}{\mathrm{Var}_{X,Y}[XY]} \simeq -1.6 \cdot 10^{-5}. \tag{3.103}$$

Example 3.11.

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

$$\Omega_{\textit{child 1}} = \{\text{👦}, \text{👧}\}, \quad \Omega_{\textit{child 2}} = \{\text{👦}, \text{👧}\}, \tag{3.104}$$

and define

$$\Omega = \Omega_{\textit{child 1}} \times \Omega_{\textit{child 2}} = \{(\text{👦},\text{👦}), (\text{👦},\text{👧}), (\text{👧},\text{👦}), (\text{👧},\text{👧})\}. \tag{3.105}$$

Define the random variables

$$B\colon \Omega \to \{0,1,2\}, \quad G\colon \Omega \to \{0,1,2\}, \tag{3.106}$$

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 3.23,

$$p_{B,G}(b,g) = \mathbb{P}_{B,G}(\{(b,g)\}) = \mathbb{P}(\{\omega \in \Omega | B(\omega) = b,\, G(\omega) = g\}). \tag{3.107}$$

For instance,

$$p_{B,G}(1,1) = \mathbb{P}(\{(\text{👦},\text{👧}), (\text{👧},\text{👦})\}) = \frac{1}{2}. \tag{3.108}$$

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

$$A = \{(b,g) \in \Omega_B \times \Omega_G \mid b \geq 1\}. \tag{3.109}$$

Using Definition 3.17 the conditional probability of exactly one girl given at least one boy is (see Remark 3.14)

$$\mathbb{P}_{B,G}(\{(1,1)\} \mid A) = \frac{\mathbb{P}_{B,G}(\{(1,1)\} \cap A)}{\mathbb{P}_{B,G}(A)}. \tag{3.110}$$

From the PMF,

$$\mathbb{P}_{B,G}(\{(1,1)\} \cap A) = p_{B,G}(1,1) = \frac{1}{2}, \tag{3.111}$$

and from Theorem 3.3

$$\mathbb{P}_{B,G}(A) = \sum_{g} \sum_{b \geq 1} p_{B,G}(b,g) = p_{B,G}(1,1) + p_{B,G}(2,0) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}. \tag{3.112}$$

Hence,

$$\mathbb{P}_{B,G}(\{(1,1)\} \mid A) = \frac{2}{3}. \tag{3.113}$$

Remark 3.14 (PMF vs. image measure on events)**.** *Let B and G be discrete random variables with joint PMF $p_{B,G}$ and image measure $\mathbb{P}_{B,G}$. From Definition 3.23, the PMF is defined only for single points $(b,g) \in \Omega_B \times \Omega_G$:*

$$p_{B,G}(b,g) = \mathbb{P}_{B,G}(\{(b,g)\}). \tag{3.114}$$

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

$$\mathbb{P}_{B,G}(\{(b,g)\} \cap A) \tag{3.115}$$

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

$$p_{B,G}(b, g \cap A) \tag{3.116}$$

(or similarly) is not valid, because $p_{B,G}$ 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

$$\mathbb{P}(E) = 0.01. \tag{3.117}$$

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

$$\mathbb{P}(E \mid \text{innocent}) = 0.01 \Rightarrow \mathbb{P}(\text{guilty} \mid E) = 0.99. \tag{3.118}$$

To investigate this claim, use Theorem 3.2 to write

$$\begin{aligned} \mathbb{P}(E \mid \text{innocent}) &= \frac{\mathbb{P}(E \cap \text{innocent})}{\mathbb{P}(\text{innocent})} \\ &= \frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})}\mathbb{P}(E). \end{aligned} \tag{3.119}$$

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,*

$$\frac{\mathbb{P}(\text{innocent} \mid E)}{\mathbb{P}(\text{innocent})} = \frac{\frac{M-1}{M}}{\frac{N-1}{N}}, \tag{3.120}$$

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 3.17, as follows*

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(\text{guilty} \cap E) + \mathbb{P}(\text{innocent} \cap E)}{\mathbb{P}(E)}. \tag{3.121}$$

Innocent and guilty are complementary events that form a partition of the sample space, meaning (Theorem 3.3)

$$\mathbb{P}(\text{guilty} \cap E) + \mathbb{P}(\text{innocent} \cap E) = \mathbb{P}(E), \tag{3.122}$$

and thereby

$$\mathbb{P}(\text{guilty} \mid E) + \mathbb{P}(\text{innocent} \mid E) = 1. \tag{3.123}$$

This means that if $\mathbb{P}(\text{guilty} \mid E) = 0.99$*, then* $\mathbb{P}(\text{innocent} \mid E) = 0.01$*, and from Theorem 3.2,*

$$\mathbb{P}(\text{innocent} \mid E) = \frac{\mathbb{P}(E \mid \text{innocent})\,\mathbb{P}(\text{innocent})}{\mathbb{P}(E)} \tag{3.124}$$

From Equation 3.124, it is clear that in general

$$\mathbb{P}(E \mid \text{innocent}) \neq \mathbb{P}(\text{innocent} \mid E), \tag{3.125}$$

and so even if $\mathbb{P}(E \mid \text{innocent}) = 0.01$*, the prosecutor's claim (Equation 3.118) 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 3.2 and Theorem 3.3

$$\begin{aligned}\mathbb{P}(s|p) &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p)} \\ &= \frac{\mathbb{P}(p|s)\mathbb{P}(s)}{\mathbb{P}(p|s)\mathbb{P}(s)+\mathbb{P}(p|h)\mathbb{P}(h)}\end{aligned} \tag{3.126}$$

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*

$$\mathbb{P}(s|p) \simeq 0.0098. \tag{3.127}$$

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

$$z^* = \underset{z}{\operatorname{argmax}}(\mathbb{P}(z|o_3 \cap c_1)). \tag{3.128}$$

Since the host cannot open the door containing the prize,

$$\mathbb{P}(z_3|o_3 \cap c_1) = 0 \tag{3.129}$$

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 3.2

$$\mathbb{P}(z_1|o_3 \cap c_1) = \frac{\mathbb{P}(o_3|c_1 \cap z_1)\mathbb{P}(c_1 \cap z_1)}{\mathbb{P}(o_3 \cap c_1)} \tag{3.130}$$

where from Theorem 3.3

$$\begin{aligned}\mathbb{P}(o_3 \cap c_1) &= \sum_i \mathbb{P}(o_3 \cap c_1 \cap z_i) \\ &= \mathbb{P}(o_3 \cap c_1 \cap z_1) + \mathbb{P}(o_3 \cap c_1 \cap z_2) + \mathbb{P}(o_3 \cap c_1 \cap z_3) \\ &= \mathbb{P}(o_3|c_1 \cap z_1)\mathbb{P}(c_1 \cap z_1) + \mathbb{P}(o_3|c_1 \cap z_2)\mathbb{P}(c_1 \cap z_2) \\ &\quad + \mathbb{P}(o_3|c_1 \cap z_3)\mathbb{P}(c_1 \cap z_3).\end{aligned} \tag{3.131}$$

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

$$\begin{aligned}\mathbb{P}(z_1|o_3 \cap c_1) &= \frac{\mathbb{P}(o_3|c_1 \cap z_1)}{\sum_i \mathbb{P}(o_3|c_1 \cap z_i)} \\ &= \frac{1}{3}.\end{aligned} \tag{3.132}$$

Similarly

$$\begin{aligned}\mathbb{P}(z_2|o_3 \cap c_1) &= \frac{\mathbb{P}(o_3|c_1 \cap z_2)}{\sum_i \mathbb{P}(o_3|c_1 \cap z_i)} \\ &= \frac{2}{3}.\end{aligned} \tag{3.133}$$

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,

$$\text{Corr}_{X,Y}[X, Y] = \text{Corr}_X[X, X^2]. \tag{3.134}$$

Using Definition 3.26 and Definition 3.25

$$\begin{aligned} \text{Corr}_X[X, X^2] &= \frac{\text{Cov}_X[X, X^2]}{\sqrt{\text{Var}_X[X]\,\text{Var}_X[X^2]}} \\ &= \frac{\mathbb{E}_X[X^3] - \mathbb{E}_X[X]\mathbb{E}_X[X^2]}{\sqrt{\text{Var}_X[X]\,\text{Var}_X[X^2]}} \end{aligned} \tag{3.135}$$

In this case for the nominator

$$\begin{aligned} \text{Cov}_X[X, X^2] &= \int_{\Omega_X} x^3 p_X(x)\mathrm{d}x - \int_{\Omega_X} x' p_X(x')\mathrm{d}x' \int_{\Omega_X} x''^2 p_X(x'')\mathrm{d}x'' \\ &= \frac{1}{b-a}\int_a^b x^3 \mathrm{d}x - \frac{1}{(b-a)^2}\int_a^b x'\mathrm{d}x' \int_a^b x''^2 \mathrm{d}x'' \\ &= \frac{1}{12}(a-b)^2(a+b) \\ &= 0 \end{aligned} \tag{3.136}$$

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

$$\begin{aligned} \text{Var}_X[X]\,\text{Var}_X[X^2] &= \left(\mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2\right)\left(\mathbb{E}_X[X^4] - \mathbb{E}_X[X^2]^2\right) \\ &= \frac{1}{540}(b-a)^4(4a^2 + 7ab + 4b^2) \\ &\neq 0. \end{aligned} \tag{3.137}$$

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

Example 3.16.

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

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

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

$$\begin{aligned} \mathbb{E}_Y[Y] &= \mathbb{E}_{W,X}[WX] \\ &= \mathbb{E}_W[W]\cancelto{0}{\mathbb{E}_X[X]} \\ &= 0. \end{aligned} \tag{3.138}$$

   The variance

$$\begin{aligned} \text{Var}_Y[Y] &= \mathbb{E}_Y[Y^2] - \cancelto{0}{\mathbb{E}_Y[Y]^2} \\ &= \mathbb{E}_{W,X}[W^2X^2] \\ &= \mathbb{E}_W[W^2]\mathbb{E}_X[X^2] \\ &= \mathbb{E}_W[W^2]\,\text{Var}_X[X] \end{aligned} \tag{3.139}$$

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

$$\begin{aligned} \mathbb{E}_W[W^2] &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 p_W(w_i) \\ &= \frac{1}{2}[(-1)^2\frac{1}{2} + 1^2\frac{1}{2}] \\ &= 1 \end{aligned} \tag{3.140}$$

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

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

$$\begin{aligned} \text{Cov}_{X,Y}[X, Y] &= \text{Cov}_{X,W}[X, WX] \\ &= \mathbb{E}_{X,W}[WX^2] - \mathbb{E}_X[X]\mathbb{E}_{X,W}[WX] \\ &= \mathbb{E}_W[W]\mathbb{E}_X[X^2] - \mathbb{E}_W[W]\mathbb{E}_X[X]^2 \\ &= \mathbb{E}_W[W]\,\text{Var}_X[X] \\ &= 0 \end{aligned} \tag{3.141}$$

where for the last equality it has been used that

$$\begin{aligned}\mathbb{E}_W[W] &= \frac{1}{n}\sum_{i=1}^{n} w_i p_W(w_i) \\ &= \frac{1}{2}[(-1)\frac{1}{2} + 1\frac{1}{2}] \\ &= 0\end{aligned} \tag{3.142}$$

Example 3.17.

According to Definition 3.22 the variance is defined as positive definite. This means

$$\begin{aligned}0 &\leq \mathrm{Var}_{X,Y}\left[\frac{X}{\sqrt{\mathrm{Var}_X[X]}} \pm \frac{Y}{\sqrt{\mathrm{Var}_Y[Y]}}\right] \\ &= \frac{\mathrm{Var}_X[X]}{\mathrm{Var}_X[X]} + \frac{\mathrm{Var}_Y[Y]}{\mathrm{Var}_Y[Y]} \pm \frac{2}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_Y[Y]}}\,\mathrm{Cov}_{X,Y}[X,Y] \\ &= 2 \pm 2\,\mathrm{Corr}_{X,Y}[X,Y].\end{aligned} \tag{3.143}$$

From Equation 3.143 the result follows

$$-1 \leq \mathrm{Corr}_{X,Y}[X,Y] \leq 1. \tag{3.144}$$

Example 3.18.

Let $X : \Omega \to \Omega_X$ be a continuous random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and suppose $Y = aX + b$ given parameters $a > 0$ and b. Since Y is uniquely determined by X,

$$\mathrm{Corr}_{X,Y}[X,Y] = \mathrm{Corr}_X[X, aX+b]. \tag{3.145}$$

Using Definition 3.26 and Definition 3.25

$$\mathrm{Corr}_X[X, aX+b] = \frac{\mathrm{Cov}_X[X, aX+b]}{\sqrt{\mathrm{Var}_X[X]\,\mathrm{Var}_X[aX+b]}} \tag{3.146}$$

with

$$\begin{aligned}\mathrm{Cov}_X[X, aX+b] &= \mathbb{E}_X[X(aX+b)] - \mathbb{E}_X[X]\mathbb{E}_X[aX+b] \\ &= a\mathbb{E}_X[X^2] + b\mathbb{E}_X[X] - a\mathbb{E}_X[X]^2 - b\mathbb{E}_X[X] \\ &= a\,\mathrm{Var}_X[X]\end{aligned} \tag{3.147}$$

and

$$\begin{aligned}\operatorname{Var}_X[aX+b] &= a^2 \operatorname{Var}_X[X] + \cancelto{0}{\operatorname{Var}_X[b]} + \cancelto{0}{2\operatorname{Cov}_{X,X}[aX,b]} \\ &= a^2 \operatorname{Var}_X[X].\end{aligned} \tag{3.148}$$

Combining Equation 3.146, Equation 3.147 and Equation 3.148 yields

$$\begin{aligned}\operatorname{Corr}_X[X,aX+b] &= \frac{a \operatorname{Var}_X[X]}{\sqrt{a^2 \operatorname{Var}_X[X] \operatorname{Var}_X[X]}} \\ &= \frac{a}{|a|}.\end{aligned} \tag{3.149}$$

CHAPTER 4

Framing of Statistics

In this book, statistics is framed as a game against Nature, following conventions from decision theory [2]. In this game, two players interact under uncertainty: the Robot, who seeks to make optimal decisions, and Nature, who determines the state of the world.

Definition 4.1 (Robot)**.** *The Robot is the primary decision maker in the statistical game. It selects actions or decisions based on available information with the aim of achieving an optimal outcome under uncertainty.*

Definition 4.2 (Nature)**.** *Nature is an opposing and unpredictable player that determines the true state of the world and thereby influences the outcomes of the statistical experiment. It represents the inherent uncertainty of the environment and the data-generating process.*

Definition 4.3 (Statistical Game)**.** *The interaction between the Robot and Nature is formalized as a statistical game under uncertainty. Let*

$$(\Omega, \mathcal{F}, \mathbb{P}) \tag{4.1}$$

be a probability space. Consider $n+1$ pairs of random variables

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \colon \Omega \to \Omega_X \times \Omega_Y, \tag{4.2}$$

where (X_i, Y_i) for $i = 1, \ldots, n$ correspond to past observations collected as data, and (X_{n+1}, Y_{n+1}) corresponds to a new observation X_{n+1} with unknown outcome Y_{n+1} to be predicted. Define the dataset as the collection of past realizations

$$D = \{(x_i, y_i)\}_{i=1}^{n} \in (\Omega_X \times \Omega_Y)^n, \tag{4.3}$$

where $x_i \in \Omega_X$ and $y_i \in \Omega_Y$ are realizations of X_i and Y_i, respectively. Each pair (x_i, y_i) corresponds to one past round of the game. The image measures of the random variables are

$$\begin{aligned} \mathbb{P}_{X_{1:n+1}} &= \mathbb{P} \circ (X_1, \ldots, X_{n+1})^{-1}, \\ \mathbb{P}_{Y_{1:n+1}} &= \mathbb{P} \circ (Y_1, \ldots, Y_{n+1})^{-1}, \\ \mathbb{P}_{(X,Y)_{1:n+1}} &= \mathbb{P} \circ ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))^{-1}, \end{aligned} \tag{4.4}$$

defined on the measurable spaces

$$(\Omega_X^{n+1}, \mathcal{F}_X^{n+1}), \quad (\Omega_Y^{n+1}, \mathcal{F}_Y^{n+1}), \quad ((\Omega_X \times \Omega_Y)^{n+1}, (\mathcal{F}_X \otimes \mathcal{F}_Y)^{n+1}). \tag{4.5}$$

Suppose that the joint image measure $\mathbb{P}_{(X,Y)_{1:n+1}}$ *depends on an unknown parameter* $w \in \Omega_W$*, where* Ω_W *is the parameter space. Let*

$$\mathcal{P}' = \{\mathbb{P}^w_{(X,Y)_{1:n+1}} \mid w \in \Omega_W\}, \tag{4.6}$$

denote a parametric family of joint image measures on $(\Omega_X \times \Omega_Y)^{n+1}$*. The statistical model of the game is then defined as the triple*

$$((\Omega_X \times \Omega_Y)^{n+1}, (\mathcal{F}_X \otimes \mathcal{F}_Y)^{n+1}, \mathcal{P}'). \tag{4.7}$$

In the game, the Robot selects an action $u_{n+1} \in \Omega_U$ *according to a decision rule*

$$U \colon \Omega_X \times (\Omega_X \times \Omega_Y)^n \to \Omega_U, \tag{4.8}$$

mapping the new observation x_{n+1} *and past dataset* D *to an action. The Robot incurs a penalty determined by a cost function*

$$C \colon \Omega_U \times \Omega_Y \to \mathbb{R}, \tag{4.9}$$

which assigns a numerical loss to each pair $(u, y) \in \Omega_U \times \Omega_Y$*. The Robot's goal is to minimize the expected cost [1]*

$$\mathbb{E}^{\mathcal{P}'}_{(X,Y)_{1:\,n+1}}[C(U(x_{n+1}, D), Y_{n+1})]. \tag{4.10}$$

The superscript $\mathcal{P}'$*, denotes the object is taken with respect to a member or mixture of the parametric family. It creates a unified notation for different statistical paradigms (see Remark 4.2 and Remark 6.2).*

Remark 4.1 (Superscript notation)**.** *The superscript* w *in* $\mathbb{P}^w_Z$ *serves as an index identifying a particular member of a family of probability measures or densities. It indicates dependence on an underlying parameter without specifying whether this parameter is fixed, unknown, or random. This convention provides a generic notation that remains valid across different interpretations of statistical models.*

Remark 4.2 (Notation of Statistics)**.** *In accordance with Definition 4.3, the expected cost (Equation 4.10) can be written as follows*

$$\begin{aligned}&\mathbb{E}^{\mathcal{P}'}_{(X,Y)_{1:\,n+1}}\big[C(U(x_{n+1}, D), Y_{n+1})\big] \\ &\quad = \int_{(\Omega_X \times \Omega_Y)^{n+1}} C\big(U(x_{n+1}, D), y_{n+1}\big)\, \mathrm{d}\mathbb{P}^{\mathcal{P}'}_{(X,Y)_{1:\,n+1}}(D, x_{n+1}, y_{n+1}).\end{aligned}$$

(4.11)

Equation 4.11 can be further decomposed by specifying the cost function or decomposition of the joint image measure via Theorem 3.1, Theorem 3.2 or Theorem 3.3. In this case, the notation of probability theory becomes cumbersome and for this reason, statistical notation is typically relaxed – when the context is reasonably clear – by omitting distribution subscripts, superscripts and integral domains. In this notation, Equation 4.11 becomes

$$\begin{aligned}\mathbb{E}\big[\, C(U(x_{n+1}, D), Y_{n+1})\big] &= \int C\big(U(x_{n+1}, D), y_{n+1}\big)\mathrm{d}\mathbb{P}(D, x_{n+1}, y_{n+1}) \\ &= \int C\big(U(\tilde{D}), y_{n+1}\big)\mathrm{d}\mathbb{P}(\tilde{D}, y_{n+1}),\end{aligned} \tag{4.12}$$

where $\tilde{D} = \{x_{n+1}, D\}$ is used to further compress the notation.

Example 4.1.

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

$$\Omega_U = \{\text{"}\textit{bring umbrella}\text{"}, \text{"}\textit{don't bring umbrella}\text{"}\}. \tag{4.13}$$

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

$$\Omega_Y = \{\text{"}\textit{rain}\text{"}, \text{"}\textit{no rain}\text{"}\}, \tag{4.14}$$

so the Robot's task is to estimate Nature's decision regarding rain later and either bring the umbrella or not.

Theorem 4.1 (Optimal Decision Rule)**.** *In accordance with Definition 4.3, the optimal decision rule for the Robot is*

$$\begin{aligned}U^*(\tilde{D}) &= \underset{U(\tilde{D})}{\operatorname{argmin}}\, \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] \\ &= \underset{U(\tilde{D})}{\operatorname{argmin}} \int C(U(\tilde{D}), y_{n+1}) p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1}.\end{aligned} \tag{4.15}$$

Proof. From Definition 4.3

$$U^*(\tilde{D}) = \underset{U(\tilde{D})}{\operatorname{argmin}}\, \mathbb{E}[C(U(\tilde{D}), Y_{n+1})], \tag{4.16}$$

where the expected cost (Equation 4.10) can be written

$$\mathbb{E}\big[\, C(U(\tilde{D}), Y_{n+1})\big] = \int C\big(U(\tilde{D}), y_{n+1}\big)\mathrm{d}\mathbb{P}(\tilde{D}, y_{n+1}). \tag{4.17}$$

From Theorem 3.7

$$\mathbb{E}[C(U(\tilde{D}), Y_{n+1})] = \mathbb{E}[\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]]. \tag{4.18}$$

Using Equation 4.18 in Equation 4.16

$$\begin{aligned} U^*(\tilde{D}) &= \underset{U(\tilde{D})}{\operatorname{argmin}} \mathbb{E}[\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]] \\ &= \underset{U(\tilde{D})}{\operatorname{argmin}} \int p(\tilde{D})\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]\mathrm{d}\tilde{D}. \end{aligned} \tag{4.19}$$

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

$$\begin{aligned} U^*(\tilde{D}) &= \underset{U(\tilde{D})}{\operatorname{argmin}} \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] \\ &= \underset{U(\tilde{D})}{\operatorname{argmin}} \int C(U(\tilde{D}), y_{n+1})p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1}. \end{aligned} \tag{4.20}$$

□

Example 4.2.

In general the random variables X_i 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 $\{x_{n+1}, D_x\} = \emptyset$ and consequently

$$\begin{aligned} \tilde{D} &= \{y_i\}_{i=1}^n \\ &\equiv D_y. \end{aligned} \tag{4.21}$$

In this case, the Robot is forced to model the decisions of Nature with a probability distribution with associated parameters without observations. From Equation 4.15 the optimal action for the Robot can be written

$$U^*(D_y) = \underset{U(D_y)}{\operatorname{argmin}} \mathbb{E}[C(U(D_y), Y_{n+1})|D_y] \tag{4.22}$$

4.1 ASSIGNING A COST FUNCTION

The cost function (Definition 4.3) 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.

4.1.1 *Continuous Action Space*

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

Definition 4.4 (Linear Cost Function)**.** *The linear cost function is defined as follows*

$$C(U(\tilde{D}), y_{n+1}) \equiv |U(\tilde{D}) - y_{n+1}|. \tag{4.23}$$

Theorem 4.2 (Median Decision Rule)**.** *The cost function of Definition 4.4 leads to the median decision rule*

$$\int_{-\infty}^{U^*(\tilde{D})} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} = \frac{1}{2}. \tag{4.24}$$

Proof. The expected cost used in Theorem 4.1

$$\begin{aligned}\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] &= \int_{-\infty}^{\infty} |U(\tilde{D}) - y_{n+1}| p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} \\ &= \int_{-\infty}^{U(\tilde{D})} (y_{n+1} - U(\tilde{D}))p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} \\ &\quad + \int_{U(\tilde{D})}^{\infty} (U(\tilde{D}) - y_{n+1})p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1}.\end{aligned} \tag{4.25}$$

The decision rule that minimize the cost can be found via

$$\begin{aligned}0 &= \frac{\partial \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]}{\partial U(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}) + \int_{-\infty}^{U^*(\tilde{D})} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} \\ &\quad + (U^*(\tilde{D}) - U^*(\tilde{D}))p(U^*(\tilde{D})|\tilde{D}) - \int_{U^*(\tilde{D})}^{\infty} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1}.\end{aligned} \tag{4.26}$$

This leads to

$$\begin{aligned}\int_{-\infty}^{U^*(\tilde{D})} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} &= \int_{U^*(\tilde{D})}^{\infty} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} \\ &= 1 - \int_{-\infty}^{U^*(\tilde{D})} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} \\ &\Downarrow \\ \int_{-\infty}^{U^*(\tilde{D})} p(y_{n+1}|\tilde{D})\mathrm{d}y_{n+1} &= \frac{1}{2}.\end{aligned} \tag{4.27}$$

□

Definition 4.5 (Quadratic Cost Function)**.** *The quadratic cost function is defined as*

$$C(U(\tilde{D}), s) \equiv (U(\tilde{D}) - s)^2. \tag{4.28}$$

Theorem 4.3 (Expectation Decision Rule)**.** *The cost function of Definition 4.5 leads to the expectation decision rule*

$$U^*(\tilde{D}) = \mathbb{E}[Y_{n+1}|\tilde{D}]. \tag{4.29}$$

Proof. From Theorem 4.1

$$\begin{aligned}
\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] &= \int (U(\tilde{D}) - y_{n+1})^2 p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1} \\
&\Downarrow \\
\frac{\partial \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]}{\partial U(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(x)} &= 2U^*(\tilde{D}) - 2\int y_{n+1} p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1} \\
&= 0 \\
&\Downarrow \\
U^*(\tilde{D}) &= \int y_{n+1} p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1} \\
&= \mathbb{E}[Y_{n+1}|\tilde{D}].
\end{aligned} \tag{4.30}$$

□

Definition 4.6 (0-1 Cost Function)**.** *The 0-1 cost function is defined as follows*

$$C(U(\tilde{D}), y_{n+1}) \equiv 1 - \delta(U(\tilde{D}) - y_{n+1}). \tag{4.31}$$

Theorem 4.4 (MAP Decision Rule)**.** *The cost function of Definition 4.6 leads to the maximum a posteriori (MAP) decision rule*

$$\frac{\partial p_{Y_{n+1}|X_{n+1},(X,Y)_{1:\,n}}(U(\tilde{D})|\tilde{D})}{\partial U(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(\tilde{D})} = 0, \tag{4.32}$$

where the distribution subscript has been included since it is not obvious from the context.

Proof. From Theorem 4.1

$$\mathbb{E}[C((\tilde{D}), Y_{n+1})|\tilde{D}] = 1 - \int \delta(U(\tilde{D}) - y_{n+1}) p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1} \tag{4.33}$$

$$\begin{aligned} \frac{\partial \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]}{\partial U(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(\tilde{D})} &= -\frac{\partial p(U(\tilde{D})|\tilde{D})}{\partial U(\tilde{D})}\bigg|_{U(\tilde{D})=U^*(\tilde{D})} \\ &= 0. \end{aligned} \tag{4.34}$$

□

Example 4.3.

The median decision rule is symmetric with respect to

$$z(\tilde{D}, y_{n+1}) \equiv U(\tilde{D}) - y_{n+1}, \tag{4.35}$$

meaning underestimation ($z < 0$) and overestimation ($z > 0$) is penalized equally. This decision rule can be generalized by adopting the cost function

$$C(U(\tilde{D}), y_{n+1}) = \alpha \cdot \mathrm{swish}(z(\tilde{D}, y_{n+1}), \beta) + (1-\alpha) \cdot \mathrm{swish}(-z(\tilde{D}, y_{n+1}), \beta), \tag{4.36}$$

where

$$\mathrm{swish}(z, \beta) = \frac{z}{1 + e^{-\beta z}}. \tag{4.37}$$

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

$$\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] = \int C(U(\tilde{D}), y_{n+1}) p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1}. \tag{4.38}$$

The derivative of the cost function with respect to the decision rule can be approximated as follows

$$\begin{aligned} \frac{dC}{dU} &= \frac{dC}{dz}\frac{dz}{dU} \\ &= \left(\frac{\alpha}{1 + e^{-\beta z}} - \frac{1-\alpha}{1 + e^{\beta z}} \right. \\ &\qquad \left. + \frac{\alpha \beta e^{-\beta z} z}{(1 + e^{-\beta z})^2} + \frac{(1-\alpha)\beta e^{\beta z} z}{(1 + e^{\beta z})^2} \right) \frac{dz}{dU} \\ &= \frac{\beta z e^{\beta z} - e^{\beta z} - 1}{(1 + e^{\beta z})^2} + \alpha + \mathcal{O}(\alpha^2) \\ &\approx \alpha - \frac{1}{(1 + e^{\beta z})^2} \end{aligned} \tag{4.39}$$

leading to the derivative of the expected cost

$$\frac{d\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}]}{dU(\tilde{D})} \approx \int \left(\alpha - \frac{1}{(1+e^{\beta z(\tilde{D}, y_{n+1})})^2}\right) p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1}$$
$$= \alpha - \int \frac{1}{(1+e^{\beta z(\tilde{D}, y_{n+1})})^2} p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1}. \tag{4.40}$$

For large β, $\frac{1}{(1+e^{\beta z(\tilde{D}, y_{n+1})})^2}$ *approaches the indicator* $\mathbb{1}\{y_{n+1} > U(\tilde{D})\}$. *Hence,*

$$\int_{-\infty}^{\infty} p(y_{n+1}|\tilde{D}) \frac{1}{(1+e^{\beta z(\tilde{D}, y_{n+1})})^2} \mathrm{d}y_{n+1} \approx \int_{U(\tilde{D})}^{\infty} p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1} \tag{4.41}$$

This means the optimal decision rule can be written as follows

$$\alpha \approx \int_{U^*(\tilde{D})}^{\infty} p(y_{n+1}|\tilde{D}) \mathrm{d}y_{n+1}. \tag{4.42}$$

The optimal decision $U^*(\tilde{D})$ *is the* α*-quantile of the conditional distribution* $p(y_{n+1}|\tilde{D})$. *This rule is known as the quantile decision rule.*

4.1.2 *Discrete Action Space*

In case of a continuous action space, the conditional expected loss used in Theorem 4.1 can be written

$$\mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] = \sum_{y_{n+1} \in \Omega_Y} C(U(\tilde{D}), y_{n+1}) p(y_{n+1}|\tilde{D}), \tag{4.43}$$

where the cost function is typically represented in matrix form as follows

		Y_{n+1}		
		$y^{(1)}$	$\dots$	$y^{(\dim(\Omega_Y))}$
$U(\tilde{D})$	$u^{(1)}$	$C(u^{(1)}, y^{(1)})$	$\dots$	$C(u^{(1)}, y^{(\dim(\Omega_Y))})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$u^{(\dim(\Omega_U))}$	$C(u^{(\dim(\Omega_U))}, y^{(1)})$	$\dots$	$C(u^{(\dim(\Omega_U))}, y^{(\dim(\Omega_Y))})$

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

Example 4.4.
With reference to Example 4.1, the possible states of Nature are $y^{(1)} =$ "*rain*" *and* $y^{(2)} =$ "*no rain*", *whereas each observed outcome* y_i *in the dataset*

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\} \tag{4.44}$$

takes a value in $\{y^{(1)}, y^{(2)}\}$. *For instance, one possible dataset realization could be* $y_1 = y^{(1)}$, $y_2 = y^{(1)}$, *and* $y_3 = y^{(2)}$.

Example 4.5.
Consider a binary classification problem with action space $\Omega_U = \{u^{(1)}, u^{(2)}\}$ *and Nature's state space* $\Omega_Y = \{y^{(1)}, y^{(2)}\}$, *where* $u^{(1)}$ *corresponds to predicting class* $y^{(1)}$ *and* $u^{(2)}$ *to predicting class* $y^{(2)}$. *Let*

$$D = \{(x_i, y_i)\}_{i=1}^{n} \tag{4.45}$$

denote the data, where $y_i \in \Omega_Y$ *are observed realizations of Nature's states. Let* $U(x_{n+1}, D)$ *be a classifier based on the probability*

$$p_{Y_{n+1}|X_{n+1},(X,Y)_{1:\,n}}(y_{n+1}|x_{n+1}, D). \tag{4.46}$$

Define a threshold $k \in [0, 1]$ *and the decision rule*

$$U_k(x_{n+1}, D) = \begin{cases} u^{(1)}, & p_{Y_{n+1}|X_{n+1},(X,Y)_{1:\,n}}(y^{(2)}|x_{n+1}, D) < k, \\ u^{(2)}, & p_{Y_{n+1}|X_{n+1},(X,Y)_{1:\,n}}(y^{(2)}|x_{n+1}, D) \geq k. \end{cases} \tag{4.47}$$

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

$U_k(x_{n+1}, D)$		Y_{n+1} $y^{(1)}$	$y^{(2)}$
	$u^{(1)}$	*TP(k)*	*FP(k)*
	$u^{(2)}$	*FN(k)*	*TN(k)*

and standard performance measures are defined as

$$\mathrm{TPR(k)} = \frac{\mathrm{TP(k)}}{\mathrm{TP(k)} + \mathrm{FN(k)}}, \tag{4.48}$$

$$\mathrm{FPR(k)} = \frac{\mathrm{FP(k)}}{\mathrm{FP(k)} + \mathrm{TN(k)}}, \tag{4.49}$$

$$\mathrm{Accuracy(k)} = \frac{\mathrm{TP(k)} + \mathrm{TN(k)}}{\mathrm{TP(k)} + \mathrm{TN(k)} + \mathrm{FP(k)} + \mathrm{FN(k)}}. \tag{4.50}$$

Example 4.6.
Consider a binary classification problem with action space $\Omega_U = \{u^{(1)}, u^{(2)}\}$ *and Nature's state space* $\Omega_Y = \{y^{(1)}, y^{(2)}\}$*, where* $u^{(1)}$ *corresponds to predicting class* $y^{(1)}$ *and* $u^{(2)}$ *to predicting class* $y^{(2)}$*. Let*

$$D = \{(x_i, y_i)\}_{i=1}^{n} \tag{4.51}$$

denote the data, where $y_i \in \Omega_Y$ *are observed realizations of Nature's states and let the cost function be defined by the matrix*

$$\begin{array}{cccc} & & Y_{n+1} & \\ & & y^{(1)} & y^{(2)} \\ U(\tilde{D}) & u^{(1)} & 0 & \lambda_{12} \\ & u^{(2)} & \lambda_{21} & 0 \end{array}$$

where λ_{12} *denotes the cost of predicting* $y^{(1)}$ *when the true state is* $y^{(2)}$*, and* λ_{21} *the cost of the reverse error. The decision* $U(\tilde{D})$ *is determined by minimizing the conditional expected cost (Equation 4.43)*

$$\begin{aligned} \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] &= \sum_{y_{n+1} \in \Omega_Y} C(U(\tilde{D}), y_{n+1}) p(y_{n+1}|\tilde{D}) \\ &= C(U(\tilde{D}), y^{(1)}) p(y^{(1)}|\tilde{D}) \\ &\quad + C(U(\tilde{D}), y^{(2)}) p(y^{(2)}|\tilde{D}). \end{aligned} \tag{4.52}$$

For the different possible actions

$$\begin{aligned} \mathbb{E}[C(u^{(1)}, Y_{n+1})|\tilde{D}] &= \lambda_{12} p(y^{(2)}|\tilde{D}), \\ \mathbb{E}[C(u^{(2)}, Y_{n+1})|\tilde{D}] &= \lambda_{21} p(y^{(1)}|\tilde{D}), \end{aligned} \tag{4.53}$$

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

$$\mathbb{E}[C(u^{(1)}, Y_{n+1})|\tilde{D}] < \mathbb{E}[C(u^{(1)}, Y_{n+1})|\tilde{D}]) \tag{4.54}$$

meaning

$$\begin{aligned} \lambda_{12} p(y^{(2)}|\tilde{D}) &< \lambda_{21} p(y^{(1)}|\tilde{D}) \\ &= \lambda_{21}(1 - p(y^{(2)}|\tilde{D})) \end{aligned} \tag{4.55}$$

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

$$p(y^{(2)}|\tilde{D}) < \frac{\lambda_{21}}{\lambda_{12} + \lambda_{21}} = k. \tag{4.56}$$

Equation 4.56 is equivalent to Equation 4.47 from Example 4.5. The difference between the two is that if λ_{12} and λ_{21} are specified, k is specified. In Example 4.5 k was left as a free parameter.

Example 4.7.

In many classification problems the Robot has the option of assigning x_{n+1} 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

$$C(U(\tilde{D}), y_{n+1}) = \begin{cases} 0 & \text{if correct classification } (U(\tilde{D}) = y_{n+1}) \\ \lambda_r & \text{if reject option } (U(\tilde{D}) = \text{reject}) \\ \lambda_s & \text{if wrong classification } (U(\tilde{D}) \neq y_{n+1}) \end{cases}. \tag{4.57}$$

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

$$\begin{aligned} \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] &= \sum_{y_{n+1} \in \Omega_Y} C(U(\tilde{D}), y_{n+1}) p(y_{n+1}|\tilde{D}) \\ &= \lambda_s (1 - p(U(\tilde{D})|\tilde{D})), \end{aligned} \tag{4.58}$$

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

$$\mathbb{E}[C(\text{reject}, Y_{n+1})|\tilde{D}] = \lambda_r, \tag{4.59}$$

and from Equation 4.58 and Equation 4.59 it follows that picking

$$\underset{U(\tilde{D}) \in \Omega_U \setminus \text{reject}}{\text{argmax}} \quad p(U(\tilde{D})|\tilde{D}) \tag{4.60}$$

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 4.58 and Equation 4.59 yields

$$(1 - p(U(\tilde{D})|\tilde{D}))\lambda_s < \lambda_r, \tag{4.61}$$

meaning

$$p(U(\tilde{D})|\tilde{D}) \geq 1 - \frac{\lambda_r}{\lambda_s}. \tag{4.62}$$

Qualitatively, as $\frac{\lambda_r}{\lambda_s}$ *is increased from* 0 *to* 1*, the behavior of the Robot changes smoothly. When*

$$\frac{\lambda_r}{\lambda_s} = 0, \tag{4.63}$$

rejection is rated as a successful classification – i.e., there is no cost associated with it – and thus becomes the best option unless

$$p(U(\tilde{D})|\tilde{D}) = 1, \tag{4.64}$$

corresponding to knowing the correct class with absolute certainty. In other words, in this limit rejection is optimal unless the Robot is completely certain of the correct class. Conversely, when

$$\frac{\lambda_r}{\lambda_s} = 1, \tag{4.65}$$

rejection is rated as a misclassification – i.e., $\lambda_r = \lambda_s$ *– and thus always incurs a cost. In this case, rejection is never chosen. For values of* $\frac{\lambda_r}{\lambda_s}$ *between these limits, an interpolation of interpretations applies, where rejection is preferred only when the Robot's uncertainty exceeds the corresponding threshold.*

CHAPTER 5

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 [15]; it can exploit group invariances inherent in the problem [16]; and it can ensure consistent marginalization of nuisance parameters [17].

Among these methods, the principle of maximum entropy[18] 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 [16, 19–22]. The remainder of this chapter develops the maximum entropy principle and illustrates how it can be applied to derive probability distributions from partial knowledge.

5.1 THE PRINCIPLE OF MAXIMUM ENTROPY

The principle of maximum entropy, first proposed by Jaynes [18], 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

$$Z \colon \Omega \to \Omega_Z \tag{5.1}$$

be a generic random variable from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z^w)$, where $\mathbb{P}_Z^w$ is the image measure of $\mathbb{P}$, parameterized by w (see Definition 4.3 and Remark 4.1). $\mathbb{P}_Z^w$ is a member or mixture of a parametric family of measures

$$\mathcal{P}' = \{\mathbb{P}_Z^w \mid w \in \Omega_W\}, \tag{5.2}$$

where Ω_W is the parameter space. The goal of the maximum entropy principle is to determine the probability measure $\mathbb{P}_Z^w$ that best represents the current state of knowledge, given a set of moment constraints. From definition Definition 3.23, the image measure $\mathbb{P}_Z^w$ admits a density p_Z^w with respect to the measure μ, so that for any event $B \in \mathcal{F}_Z$,

$$\mathbb{P}_Z^w(B \mid w) = \int_B p_Z^w(z \mid w) \mathrm{d}\mu(z). \tag{5.3}$$

The maximum entropy principle asserts that the density $p_Z^w(z \mid w)$ that best represents the current state of knowledge is the one that maximizes the constrained Shannon entropy [3], where w denotes the parameters of the distribution. The Shannon entropy of a probability density $p_{Z|w}$ can be written

$$H[p_Z^w] = -\int_{\Omega_Z} p_Z^w(z \mid w) \ln \frac{p_Z^w(z \mid w)}{m(z)} \mathrm{d}\mu(z), \tag{5.4}$$

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 $w_0, w_1, \ldots, w_n$ and defines the Lagrangian functional, which represents the constrained entropy:

$$\mathcal{L}[p_Z^w] = -\int_{\Omega_Z} p_Z^w(z \mid w) \left(\ln \frac{p_{Z|I}^w(z \mid w)}{m(z)} + w_0 + \sum_{j=1}^{n} w_j C_j(z) \right) \mathrm{d}\mu(z), \tag{5.5}$$

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

$$\frac{\partial}{\partial p_Z^w(z \mid w)} p_Z^w(z \mid w) \Big(\ln \frac{p_Z^w(z \mid w)}{m} + w_0 + \sum_{j=1}^{n} w_j C_j \Big) = 0, \tag{5.6}$$

which simplifies to

$$\ln \frac{p_Z^w(z \mid w)}{m(z)} + 1 + w_0 + \sum_{j=1}^{n} w_j C_j(z) = 0. \tag{5.7}$$

Hence the maximum-entropy distribution takes the exponential family form

$$\begin{aligned} p_Z^w(z \mid w) &= m(z) e^{-1-w_0-\sum_{j=1}^n w_j C_j(z)} \\ &= \frac{m(z) e^{-\sum_{j=1}^n w_j C_j(z)}}{\int_{\Omega_Z} m(z') e^{-\sum_{j=1}^n w_j C_j(z')} \mathrm{d}\mu(z')}. \end{aligned} \tag{5.8}$$

The constants w_j are determined by the imposed constraints. The resulting probability measure (Equation 5.3) defines the unique maximum-entropy probability measure consistent with the given information.

Example 5.1.

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

$$p_Z^w(z \mid w) = m(z) e^{-1-w_0-w_1 z-w_2 z^2}. \tag{5.9}$$

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

$$\begin{aligned} \int_{\Omega_Z} p_Z^w(z \mid w) \mathrm{d}z &= m e^{-1-w_0} \int_{\Omega_Z} e^{-w_1 z-w_2 z^2} \mathrm{d}z \\ &= m e^{-1-w_0} \sqrt{\frac{\pi}{w_2}} e^{\frac{w_1^2}{4w_2}} \\ &= 1. \end{aligned} \tag{5.10}$$

Defining $K^{-1} = m e^{-1-w_0}$ yields

$$\begin{aligned} p_Z^w(z \mid w) &= \frac{e^{-w_1 z-w_2 z^2}}{K} \\ &= \sqrt{\frac{w_2}{\pi}} e^{-\frac{w_1^2}{4w_2}-w_1 z-w_2 z^2}. \end{aligned} \tag{5.11}$$

Now, imposing the mean constraint

$$\begin{aligned} \int_{\Omega_Z} p_Z^w(z \mid w) \mathrm{d}z &= \frac{\int_{\Omega_Z} z e^{-w_1 z-w_2 z^2} \mathrm{d}z}{K} \\ &= -\frac{w_1}{2w_2} \\ &= \mu. \end{aligned} \tag{5.12}$$

Hereby

$$\begin{aligned} p_Z^w(z \mid w) &= \sqrt{\frac{w_2}{\pi}} e^{-\mu^2 w_2 + 2\mu w_2 z - w_2 z^2} \\ &= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{\mu - z}{\sigma}\right)^2}, \end{aligned} \tag{5.13}$$

where $\sigma \equiv \frac{1}{2w_2}$ *has been defined. Equation 5.13 can be identified as the normal distribution.*

Example 5.2.

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

$$p_Z^w(z \mid w) = m(z) e^{-1 - w_0 - w_1 \ln z - w_2 \ln(1-z)}. \tag{5.14}$$

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

$$\begin{aligned} \int_{\Omega_Z} p_Z^w(z \mid w) \mathrm{d}z &= m e^{-1-w_0} \int_{\Omega_Z} z^{-w_1} (1 - z)^{-w_2} \mathrm{d}z \\ &= m e^{-1-w_0} \frac{w(1 - w_1) w(1 - w_2)}{w(2 - w_1 - w_2)} \\ &= 1. \end{aligned} \tag{5.15}$$

Now define $\alpha \equiv 1 - w_1 \wedge \beta \equiv 1 - w_2$*. Hereby*

$$p_Z^w(z \mid \alpha, \beta) = \frac{w(\alpha + \beta)}{w(\alpha) w(\beta)} x^{\alpha - 1} (1 - x)^{\beta - 1}, \tag{5.16}$$

which is the beta distribution.

Example 5.3.

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*

$$\begin{aligned} p_Z^w(z \mid w) &= m(z) e^{-1 - w_0 - w_1 z - w_2 \ln z} \\ &= \tilde{m}(z) z^{-w_2} e^{-w_1 z} \end{aligned} \tag{5.17}$$

where $\tilde{m}(z) = m(z) e^{-1-w_0}$*. Taking a uniform measure (*$m(z) = const$*) and imposing normalization*

$$\begin{aligned} \int_{\Omega_Z} p_Z^w(z \mid w) \mathrm{d}z &= \tilde{m} \int_{\Omega_Z} z^{-w_2} e^{-w_1 z} \mathrm{d}z \\ &= 1. \end{aligned} \tag{5.18}$$

The integral is recognized as the Gamma function

$$\int_{\Omega_Z} z^{\alpha-1} e^{-\beta z} \mathrm{d}z = \frac{w(\alpha)}{\beta^\alpha} \tag{5.19}$$

with $\alpha = 1 - w_2$ *and* $\beta = w_1$*. Substituting* $\tilde{m}$*,* α*,* β *back into Equation 5.17*

$$p_Z^w(z \mid w) = \frac{\beta^\alpha}{w(\alpha)} z^{\alpha-1} e^{-\beta z}, \tag{5.20}$$

which is the Gamma distribution.

Example 5.4.

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

$$p_Z^w(z \mid w) = m(z) e^{-1-w_0-w_1 z}. \tag{5.21}$$

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

$$\begin{aligned} \int_{\Omega_Z} p_Z^w(z \mid w) \mathrm{d}z &= m e^{-1-w_0} \int_{\Omega_Z} e^{-w_1 z} \mathrm{d}z \\ &= m e^{-1-w_0} \frac{1}{w_1} \\ &= 1 \end{aligned} \tag{5.22}$$

and the mean constraint

$$\begin{aligned} \int_{\Omega_Z} z p(z \mid w) \mathrm{d}z &= \int_{\Omega_Z} z w_1 e^{-w_1 z} \mathrm{d}z \\ &= \frac{1}{w_1} \\ &= \mu. \end{aligned} \tag{5.23}$$

Combining Equation 5.22, Equation 5.23 and Equation 5.21 yields

$$p_Z^w(z \mid w) = \frac{\beta^\alpha}{w(\alpha)} z^{\alpha-1} e^{-\beta z}, \tag{5.24}$$

which is the Gamma distribution.

Example 5.5.

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

$$p_Z^w(z \mid w) = m(z) e^{-1-w_0-w_1 z}. \tag{5.25}$$

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

$$\begin{aligned}\sum_{z=0}^{1} p_Z^w(z \mid w) &= me^{-1-w_0}\left(1 + e^{-w_1}\right) \\ &= 1\end{aligned} \tag{5.26}$$

and mean constraint

$$\begin{aligned}\sum_{z=0}^{1} z p_Z^w(z \mid w) &= me^{-1-w_0}e^{-w_1} \\ &= \frac{1}{1+e^{w_1}} \\ &= \mu\end{aligned} \tag{5.27}$$

This means

$$\begin{aligned}p_Z^w(0|w) &= me^{-1-w_0} \\ &= \frac{1}{1+e^{-w_1}} \\ &= 1-\mu\end{aligned} \tag{5.28}$$

and

$$\begin{aligned}p_Z^w(1 \mid w) &= me^{-1-w_0-w_1} \\ &= \mu,\end{aligned} \tag{5.29}$$

or

$$p_Z^w(z \mid w) = \mu^z(1-\mu)^{1-z}. \tag{5.30}$$

which is the Bernoulli distribution.

Example 5.6.

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

$$p_Z^w(z \mid w) = m(z)e^{-w_0-w_1 z}. \tag{5.31}$$

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

$$\begin{aligned}\sum_{z=0}^{n} p_Z^w(z \mid w) &= \sum_{z=0}^{n} \binom{n}{z} e^{-w_0-w_1 z} \\ &= 1,\end{aligned} \tag{5.32}$$

yields

$$e^{-w_0} = (1 + e^{-w_1})^{-n}. \tag{5.33}$$

The mean constraint

$$\begin{aligned} \sum_{z=0}^{n} z p_Z^w(z \mid w) &= n \frac{e^{-w_1}}{1 + e^{-w_1}} \\ &= n\mu \end{aligned} \tag{5.34}$$

gives

$$e^{-w_1} = \frac{\mu}{1 - \mu}. \tag{5.35}$$

Finally, substituting e^{-w_0} and e^{-w_1} into $p_Z^w(z \mid w)$ gives the maximum entropy distribution

$$p_Z^w(z|w) = \binom{n}{z} \mu^z (1 - \mu)^{n-z}, \tag{5.36}$$

which is the Binomial distribution.

Example 5.7.

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

$$p_Z^w(z|w) = m(z) e^{-1-w_0-w_1 z}. \tag{5.37}$$

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

$$\begin{aligned} \sum_{z=0}^{\infty} p_Z^w(z \mid w) &= \sum_{z=0}^{\infty} \frac{e^{-1-w_0-w_1 z}}{z!} \\ &= e^{-1-w_0} \sum_{z=0}^{\infty} \frac{e^{-w_1 z}}{z!} \\ &= 1, \end{aligned} \tag{5.38}$$

Identifying the sum with the Taylor expansion

$$\sum_{z=0}^{\infty} \frac{e^{-w_1 z}}{z!} = e^{e^{-w}} \tag{5.39}$$

yields

$$e^{-1-w_0} = e^{-e^{-w_1}} \quad \Rightarrow \quad 1 + w_0 = e^{-w_1}. \tag{5.40}$$

Imposing the mean constraint

$$\begin{aligned}\sum_{z=0}^{\infty} z p_Z^w(z \mid w) &= e^{-1-w_0} \sum_{z=1}^{\infty} \frac{z e^{-w_1 z}}{z!} \\ &= e^{-1-w_0} \sum_{z=1}^{\infty} \frac{e^{-w_1 z}}{(z-1)!} \\ &= e^{-w_1} e^{-1-w_0} \sum_{y=0}^{\infty} \frac{e^{-w_1 y}}{y!} \\ &= e^{-w_1} \\ &= \mu\end{aligned} \tag{5.41}$$

where Equation 5.38 and $y = z - 1$ has been used. Combining Equation 5.37, Equation 5.40 and Equation 5.41 yield

$$p_Z^w(z \mid w) = \frac{\mu^z e^{-\mu}}{z!}. \tag{5.42}$$

which is the Poisson distribution.

CHAPTER 6

Statistical Paradigms

Axiom 6.1 (Parameter Fixedness)**.** *In the statistical game of Definition 4.3, the parameter $w \in \Omega_W$ is treated as a fixed but unknown constant. In this setting, the image measure*

$$\begin{aligned} \mathbb{P}^{\mathcal{P}'}_{(X,Y)_{1:\,n+1}} &= \mathbb{P}^{w}_{(X,Y)_{1:n+1}} \\ &= \mathbb{P}_{(X,Y)_{1:n+1}|w} \end{aligned} \tag{6.1}$$

is assumed to equal $\mathbb{P}^{w^*}_{(X,Y)_{1:n+1}}$ for some (unknown) $w^* \in \Omega_W$.

Axiom 6.2 (Parameter as a Random Variable)**.** *In the statistical game of Definition 4.3, the parameter $w \in \Omega_W$ is treated as the realization of a random variable*

$$W \colon \Omega \to \Omega_W \tag{6.2}$$

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

$$\mathbb{P}_W = \mathbb{P} \circ W^{-1}. \tag{6.3}$$

The joint image measure

$$\begin{aligned} \mathbb{P}^{\mathcal{P}'}_{(X,Y)_{1:\,n+1}} &= \mathbb{P}_{(X,Y)_{1:n+1},W} \\ &= \mathbb{P}_{(X,Y)_{1:n+1}|W}\mathbb{P}_W, \end{aligned} \tag{6.4}$$

where $\mathbb{P}_{(X,Y)_{1:n+1}|W} = \mathbb{P}^{w}_{(X,Y)_{1:n+1}}$, is defined on

$$((\Omega_X \times \Omega_Y)^{n+1} \times \Omega_W, (\mathcal{F}_X \otimes \mathcal{F}_Y)^{n+1} \otimes \mathcal{F}_W). \tag{6.5}$$

Definition 6.1 (Frequentist Statistics Paradigm)**.** *Frequentist statistics is a paradigm within the statistical game framework (Definition 4.3) that treats parameters as fixed (Axiom 6.1) and assumes an objective probability measure (Definition 3.13).*

Definition 6.2 (Bayesian Statistics Paradigm)**.** *Bayesian statistics is a paradigm within the statistical game framework (Definition 4.3) that treats parameters as realizations of a random variable (Axiom 6.2) and assumes a subjective probability measure (Definition 3.14).*

Remark 6.1 (Inference)**.** *The distinction between Bayesian and Frequentist statistics lies not in the mathematics of probability, but in the interpretation of what probability and parameters represent.*

Remark 6.2 (Frequentist and Bayesian Forms of the Expected Cost)**.** *In the paradigm of Frequentist statistics, Equation 4.10 becomes*

$$\begin{aligned}&\mathbb{E}_{(X,Y)_{1:\,n+1}|w}\big[C(U(x_{n+1},D),Y_{n+1}) \mid w\big]\\&\quad=\int_{(\Omega_X\times\Omega_Y)^{n+1}} C\big(U(x_{n+1},D),y_{n+1}\big)\\&\qquad \mathrm{d}\mathbb{P}_{(X,Y)_{1:\,n+1}|w}(D,x_{n+1},y_{n+1}\mid w),\end{aligned} \tag{6.6}$$

whereas in the paradigm of Bayesian statistics, Equation 4.10 becomes

$$\begin{aligned}&\mathbb{E}_{(X,Y)_{1:\,n+1},W}\big[C(U(x_{n+1},D),Y_{n+1})\big]\\&\quad=\int_{(\Omega_X\times\Omega_Y)^{n+1}\times\Omega_W} C\big(U(x_{n+1},D),y_{n+1}\big)\\&\qquad \mathrm{d}\mathbb{P}_{(X,Y)_{1:\,n+1},W}(D,x_{n+1},y_{n+1},w).\end{aligned} \tag{6.7}$$

6.1 BAYESIAN STATISTICS

The Bayesian paradigm (Definition 6.2) originally come from the work of Bayes [23] and Laplace [24] with much of the modern discussions and formalism created later by Finetti [25], Jeffreys [26], and Savage [27].
In the Bayesian paradigm, it is assumed that Nature's decisions can be captured by a statistical model with parameters that are modeled as realizations of random variables and a subjective probability measure. This means that the density for Nature's actions, used to determine the optimal decision rule in Theorem 4.1, can be written

$$\begin{aligned}p(y_{n+1}\mid\tilde{D}) &= \int p(w,y_{n+1}\mid\tilde{D})\mathrm{d}w\\&= \int p(y_{n+1}\mid w,\tilde{D})p(w\mid\tilde{D})\mathrm{d}w.\end{aligned} \tag{6.8}$$

Using Theorem 3.1-Theorem 3.3

$$\begin{aligned} p(w|\tilde{D}) &= p(w|D) \\ &= \frac{p(D_y|w, D_x)p(w)}{p(D_y|D_x)}, \end{aligned} \tag{6.9}$$

where $D_y = \{y_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$, Axiom 6.3 has been used for the first and second equality and $p(D_y|D_x)$ can be expanded via marginalization.

Axiom 6.3 (Relevance of Observations)**.** *The Robot's observations are relevant for estimating Nature's model only when they map to known actions of Nature.*

$p(w)$ is the Robot's prior belief about w. $p(D_s|w, D_x)$ is the likelihood of the past observations of Nature's actions, and $p(w|D)$ called the posterior distribution represent the belief of the Robot after seeing data. The prior depends on parameters that must be specified and cannot be learned from data since it reflects the Robot's belief before observing data.

6.1.1 *Bayesian Regression*

Regression involves the Robot building a model,

$$f : \Omega_\theta \times \Omega_X \to \Omega_Y, \tag{6.10}$$

with associated parameters $\theta \in \Omega_\Theta \subseteq \Omega_W$, that estimates Nature's actions $y_{n+1} \in \Omega_Y = \mathbb{R}$ based on observed data $x_{n+1} \in \Omega_X$. 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 4.5 will be considered to review the subject. According to Theorem 4.3 the best action for the Robot can be written

$$U^*(x_{n+1}, D) = \int y_{n+1} p(y_{n+1}|x_{n+1}, D) \mathrm{d}y_{n+1}. \tag{6.11}$$

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

$$p(y_{n+1}|x_{n+1}, w) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(\theta, x_{n+1}) - y_{n+1})^2}, \tag{6.12}$$

where $w = \{\theta, \xi, \dots\}$ denotes the collection of parameters. Using Equation 6.12 and marginalizing (Theorem 3.5 and Remark 3.5) over ξ, θ

$$\begin{aligned} p(y_{n+1}|x_{n+1}, D) &= \int p(y_{n+1}, \theta, \xi|x_{n+1}, D)\mathrm{d}\theta\mathrm{d}\xi \\ &= \int p(y_{n+1}|x_{n+1}, \theta, \xi, D)p(\theta, \xi|x_{n+1}, D)\mathrm{d}\theta\mathrm{d}\xi \\ &= \int p(y_{n+1}|x_{n+1}, \theta, \xi)p(\theta, \xi|D)\mathrm{d}\theta\mathrm{d}\xi, \end{aligned} \tag{6.13}$$

where it has been used that

$$p(y_{n+1}|\theta, \xi, x_{n+1}, D) = p(y_{n+1}|\theta, \xi, x_{n+1}) \tag{6.14}$$

since by definition f produce a $1-1$ map of the input x_{n+1} (Equation 6.12) and

$$p(\theta, \xi|x_{n+1}, D) = p(\theta, \xi|D) \tag{6.15}$$

from Axiom 6.3. Using Equation 6.13 in Equation 6.11[1]

$$\begin{aligned} U^*(x_{n+1}, D) &= \int f(\theta, x_{n+1})p(\theta, \xi|D)\mathrm{d}\theta\mathrm{d}\xi, \\ &= \mathbb{E}[f|x_{n+1}, D] \end{aligned} \tag{6.16}$$

where it has been used that

$$\begin{aligned} \mathbb{E}[Y_{n+1}|x_{n+1}, \theta, \xi] &= \int y_{n+1}p(y_{n+1}|x_{n+1}, \theta, \xi)\mathrm{d}y_{n+1} \\ &= f(\theta, x_{n+1}) \end{aligned} \tag{6.17}$$

according to Equation 6.12. Using Bayes theorem (Theorem 3.2)

$$p(\theta, \xi|D) = \frac{p(D_y|D_x, \theta, \xi)p(\theta, \xi|D_x)}{p(D_y|D_x)} \tag{6.18}$$

where from marginalization (Theorem 3.3)

$$p(D_y|D_x) = \int p(D_y|D_x, \theta, \xi)p(\theta, \xi|D_x)\mathrm{d}\theta\mathrm{d}\xi. \tag{6.19}$$

Assuming the past actions of Nature are independent and identically distributed (IID), the likelihood can be written (using equation Equation 6.12)

$$p(D_y|D_x, \theta, \xi) = \left(\frac{\xi}{2\pi}\right)^{\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{\xi}{2}(f(\theta, x_i)-y_i)^2} \tag{6.20}$$

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

From the chain rule (see Theorem 3.1) and Theorem 6.3

$$p(\theta, \xi | D_x) = p(\theta|\xi)p(\xi). \tag{6.21}$$

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

$$\begin{aligned} p(\theta|\xi) &= p(\theta) \\ &= \int p(\theta|\lambda)p(\lambda)\mathrm{d}\lambda \end{aligned} \tag{6.22}$$

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

$$p(\theta|\lambda)p(\lambda) = \prod_{q=1}^{\tilde{n}} \frac{\lambda_q^{\frac{n_q}{2}}}{(2\pi)^{\frac{n_q}{2}}} e^{-\frac{\lambda_q}{2}\sum_{l=1}^{n_q}\theta_l^2} \frac{\beta_q^{\alpha_q}}{\Gamma(\alpha_q)} \lambda_q^{\alpha_q - 1} e^{-\beta_q \lambda_q} \tag{6.23}$$

where α_q, β_q are prior parameters and $\tilde{n}$ is the number of hyper parameters. In the completely general case $\tilde{n}$ would equal the number of parameters θ, 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 $p(\xi)$ is analogous to $p(\lambda)$ – in that both are prior distributions for precision parameters – $p(\xi)$ is assumed to be a wide gamma distribution, then

$$\begin{aligned} p(\xi) &= \mathrm{Ga}(\xi|\tilde{\alpha}, \tilde{\beta}) \\ &= \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \xi^{\tilde{\alpha}-1} e^{-\tilde{\beta}\xi}. \end{aligned} \tag{6.24}$$

At this point equation Equation 6.11 is fully specified and can be approximated by obtaining samples from $p(\theta, \xi, \lambda|D)$ via Hamiltonian Monte Carlo (HMC) [29–32] (see Appendix A for a review of HMC). The centerpiece in the HMC algorithm is the Hamiltonian defined as follows [31, 32]

$$H \equiv \sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \ln p(\theta, \xi, \lambda|D) + const, \tag{6.25}$$

where

$$p(\theta, \xi|D) = \int p(\theta, \xi, \lambda|D)\mathrm{d}\lambda. \tag{6.26}$$

2 The normally distributed prior is closely related to weight decay [28], a principle conventionally used in Frequentist statistics to avoid the issue of overfitting.

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 paradigm[3]. Using Equation 6.18-Equation 6.26 yields

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} + \frac{n}{2}[\ln(2\pi) - \ln \xi] + \frac{\xi}{2} \sum_{i=1}^{n} (f(\theta, x_i) - y_i)^2 \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln \Gamma(\alpha_q) - \alpha_q \ln(\beta_q) + (1 - \alpha_q) \ln \lambda_q + \beta_q \lambda_q \\ &\qquad + \frac{n_q}{2} (\ln(2\pi) - \ln \lambda_q) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} \theta_l^2 \Bigg) \\ &+ \ln \Gamma(\tilde{\alpha}) - \tilde{\alpha} \ln \tilde{\beta} + (1 - \tilde{\alpha}) \ln \xi + \tilde{\beta} \xi + const. \end{aligned} \tag{6.27}$$

Example 6.1.

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 6.16 means $p(\theta, \xi, \lambda|D)$ is multiplied with ξ. Hence, when taking $-\ln p(\theta, \xi, \lambda|D)$ according to Equation 6.25, a $-\ln \xi$ is added to the Hamiltonian. In practice this means

$$(1 - \tilde{\alpha}) \ln \xi \in H \Rightarrow -\tilde{\alpha} \ln \xi. \tag{6.28}$$

6.1.2 *Bayesian Classification*

Classification involves the Robot building a model,

$$f : \Omega_\Theta \times \Omega_X \to \Delta^K, \tag{6.29}$$

with associated parameters $\theta \in \Omega_\Theta \subseteq \Omega_W$, that estimates Nature's actions $y_{n+1} \in \Omega_Y = \{1, \ldots, K\}$ based on observed data $x_{n+1} \in \Omega_X$. Here

$$\Delta^K = \left\{ p \in \mathbb{R}^K \,\middle|\, p_{y_{n+1}} \geq 0, \sum_{y_{n+1}=1}^{K} p_{y_{n+1}} = 1 \right\} \tag{6.30}$$

denotes the K-dimensional probability simplex, so that for $x_{n+1} \in \Omega_X$ the model output $f(\theta, x_{n+1})$ is a probability vector representing the conditional

3 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 [33–35]

distribution of the class label $y_{n+1} \in \Omega_Y$. In particular, the probability of observing class y_{n+1} given x_{n+1} and parameters θ is

$$p(y_{n+1} \mid x_{n+1}, \theta) = f_{y_{n+1}}(\theta, x_{n+1}), \tag{6.31}$$

where $f_{y_{n+1}}(\theta, x_{n+1})$ denotes the y_{n+1}-th component of $f(\theta, x_{n+1})$. By construction, these probabilities satisfy

$$\sum_{y_{n+1} \in \Omega_Y} p(y_{n+1} \mid x_{n+1}, \theta) = 1. \tag{6.32}$$

In this case, the Robot's action space is equal to Natures action space, with the possible addition of a reject option, $\Omega_U = \Omega_Y \cup \{\text{reject}\}$. 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 4.6), with the addition of a reject option at cost λ. This means

$$C(U(\tilde{D}), y_{n+1}) = 1 - \delta_{U(\tilde{D}),y_{n+1}} + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}}. \tag{6.33}$$

The optimal decision rule for the robot can the be written (Equation 4.43)

$$\begin{aligned} U^*(\tilde{D}) &= \operatorname*{argmin}_{U(\tilde{D})} \mathbb{E}[C(U(\tilde{D}), Y_{n+1})|\tilde{D}] \\ &= \operatorname*{argmin}_{U(\tilde{D})} \left(\sum_{y_{n+1} \in \Omega_Y} C(U(\tilde{D}), y_{n+1}) p(y_{n+1} \mid \tilde{D}) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right) \\ &= \operatorname*{argmin}_{U(\tilde{D})} \left(1 - p(U(\tilde{D})|\tilde{D}) + (\lambda - 1)\delta_{U(\tilde{D}),\text{reject}} \right). \end{aligned} \tag{6.34}$$

In absence of the reject option, the optimal decision rule is to pick the MAP, similar to Theorem 4.4. Using Equation 6.31 and marginalizing over θ

$$\begin{aligned} p(U(\tilde{D})|\tilde{D}) &= \int p(U(\tilde{D}), \theta \mid \tilde{D}) \mathrm{d}\theta \\ &= \int p(U(\tilde{D}) \mid \theta, \tilde{D}) p(\theta \mid \tilde{D}) \mathrm{d}\theta \\ &= \int p(U(\tilde{D}) \mid x_{n+1}, \theta) p(\theta \mid D) \mathrm{d}\theta \\ &= \int f_{U(\tilde{D})}(\theta, x_{n+1}) p(\theta \mid D) \mathrm{d}\theta \\ &= \mathbb{E}[f_{U(\tilde{D})}(\theta, x_{n+1}) \mid D], \end{aligned} \tag{6.35}$$

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

$$p(U(\tilde{D}) \mid \theta, \tilde{D}) = p(U(\tilde{D}) \mid \theta, x_{n+1}) \tag{6.36}$$

since by definition f (see Equation 6.31) produce a $1-1$ map of the input x_{x+1} and $p(\theta \mid \tilde{D}) = p(\theta \mid D)$ from Axiom 6.3. From Bayes theorem

$$p(w|D) = \frac{p(D_y|D_x, \theta)p(\theta \mid D_x)}{p(D_y \mid D_x)}, \tag{6.37}$$

where from Axiom 6.3 $p(\theta \mid D_x) = p(\theta)$. Assuming the distribution over θ is normally distributed with zero mean and a precision described by a hyperparameter, λ,

$$p(\theta) = \int p(\theta \mid \lambda)p(\lambda)\mathrm{d}\lambda. \tag{6.38}$$

where $p(\theta|\lambda)p(\lambda)$ is given by Equation 6.23. Assuming the past actions of Nature are independent and identically distributed, the likelihood can be written [36]

$$\begin{aligned} p(D_y|D_x, \theta) &= \prod_{i=1}^{n} p(y_i|x_i, \theta) \\ &= \prod_{i=1}^{n} f_{y_i}(\theta, x_i). \end{aligned} \tag{6.39}$$

At this point Equation 6.34 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

$$H \equiv \sum_q \sum_l \frac{p_l^2}{2m_l} - \ln p(\theta, \lambda|D) + const \tag{6.40}$$

where

$$p(\theta|D) = \int p(\theta, \lambda|D)\mathrm{d}\lambda. \tag{6.41}$$

Using Equation 6.35-Equation 6.39 in equation (6.40) yields the Hamiltonian

$$\begin{aligned} H = &\sum_{q=1}^{\tilde{n}} \sum_{l=1}^{n_q} \frac{p_l^2}{2m_l} - \sum_{i=1}^{n} \ln f_{y_i}(\theta, x_i) + \text{const} \\ &+ \sum_{q=1}^{\tilde{n}} \Bigg(\ln \Gamma(\alpha_q) - \alpha_q \ln \beta_q + (1-\alpha_q) \ln \lambda_q + \beta_q \lambda_q. \\ &\qquad + \frac{n_q}{2}(\ln(2\pi) - \ln \lambda_q) + \frac{\lambda_q}{2} \sum_{l=1}^{n_q} \theta_l^2 \Bigg) \end{aligned} \tag{6.42}$$

Sampling Equation 6.42 yields a set of coefficients which can be used to compute $\mathbb{E}[f_{y_{n+1}}(\theta, x_{n+1}) \mid D]$ which in turn (see Equation 6.35 and Equation 6.34) can be used to compute $U^*(\tilde{D})$.

6.1.3 *Making Inference About the Model of Nature*

In some instances, the robot is interested in inference related to the model of Nature. The observation $x_{n+1} \in \Omega_X$ by definition does not have an associated known action of Nature and thus by Axiom 6.3 is disregarded in this context. From Equation 4.15

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{Y_{n+1}|D}[C(U(D), Y_{n+1}) \mid D] \tag{6.43}$$

where $y_{n+1} \in \Omega_Y$ 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 $y^{(1)}$ and $y^{(2)}$ which represent which model that in truth fit Nature's true model best. From Equation 6.43

$$\begin{aligned} \mathbb{E}_{Y_{n+1}|D}[C(u_1, Y_{n+1})|D] &= \sum_{y_{n+1}=y^{(1)},y^{(2)}} C(u_1, y_{n+1}) p(y_{n+1}|D), \\ \mathbb{E}_{Y_{n+1}|D}[C(u_2, Y_{n+1})|D] &= \sum_{y_{n+1}=y^{(1)},y^{(2)}} C(u_2, y_{n+1}) p(y_{n+1}|D). \end{aligned} \tag{6.44}$$

Since there is no input x_{n+1} in this case, the decision rule U is fixed (i.e. it does not depend on x_{n+1}) given data D. $U^*(D) = u_1$ is picked iff

$$\mathbb{E}_{Y_{n+1}|D}[C(u_1, Y_{n+1})|D] < \mathbb{E}_{Y_{n+1}|D}[C(u_2, Y_{n+1})|D], \tag{6.45}$$

meaning

$$\frac{p(y^{(1)}|D)}{p(y^{(2)}|D)} > \frac{C(u_1, y^{(2)}) - C(u_2, y^{(2)})}{C(u_2, y^{(1)}) - C(u_1, y^{(1)})}. \tag{6.46}$$

The ratio $\frac{p(y^{(1)}|D)}{p(y^{(2)}|D)}$ is referred to as the posterior ratio. Using Bayes theorem it can be re-written as follows

$$\begin{aligned} \text{posterior ratio} &= \frac{p(y^{(1)}|D)}{p(y^{(2)}|D)} \\ &= \frac{p(D_y|y^{(1)}, D_x)p(y^{(1)})}{p(D_s|y^{(2)}2, D_x)p(y^{(2)})}, \end{aligned} \tag{6.47}$$

where for the second equality it has been used that the normalization $p(D)$ cancels out between the denominator and nominator and Axiom 6.3 has been employed. Given there is no a priori bias towards any model,

$$p(y^{(1)}) = p(y^{(2)}) \tag{6.48}$$

meaning

$$\text{posterior ratio} = \frac{p(D_y|y^{(1)}, D_x)}{p(D_y|y^{(2)}, D_x)}. \tag{6.49}$$

$p(D_y|y^{(1)}, D_x)$ and $p(D_y|y^{(2)}, D_x)$ can then be expanded via marginalization, the chain rule and Bayes theorem until they can be evaluated either analytically or numerically. Equation 6.49 is referred to as Bayes factor and as a rule of thumb

Definition 6.3 (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 4.15)

$$U^*(D) = \arg\min_{U(D)} \mathbb{E}_{W|D}[C(U(D), W)|D], \tag{6.50}$$

with

$$\mathbb{E}_{W|D}[C(U(D), W)|D] = \int C(U(D), w)p(w|D)\mathrm{d}w. \tag{6.51}$$

At this point, the Robot can select a cost function like in Section 4.1 and proceed by expanding $p(w|D)$ similarly to Equation 6.9. Picking the quadratic cost (Definition 4.5) yields

$$U^*(D) = \mathbb{E}_{W|D}[W|D] \tag{6.52}$$

$p(w|D)$ in Equation 6.52 can be expanded as shown in Equation 6.9.

Example 6.2.

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

$$p(w_B > w_A|D) = \int_0^1 \int_{w_A}^1 p(w_A, w_B|D)\mathrm{d}w_A\mathrm{d}w_B. \tag{6.53}$$

Assuming the costumer sets are independent

$$\begin{aligned} p(w_A, w_B|D) &= p(w_B|w_A, D)p(w_A|D) \\ &= p(w_B|D_A)p(w_A|D_A), \end{aligned} \tag{6.54}$$

with

$$p(w_i|D_i) = \frac{p(D_i|w_i)p(w_i)}{p(D_i)}. \tag{6.55}$$

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

$$p(w_i|D_i) = \frac{w_i^{\alpha_i - 1}(1 - w_i)^{\beta_i - 1}}{B(\alpha_i, \beta_i)}, \tag{6.56}$$

where $\alpha_i \equiv \alpha + y_i$, $\beta_i \equiv \beta + f_i$ and y_i / f_i denotes the successes/failure, respectively, registered in the two sets of costumers. Evaluating Equation 6.53 yields

$$p(w_B > w_A|D) = \sum_{j=0}^{\alpha_B - 1} \frac{B(\alpha_A + j, \beta_A + \beta_B)}{(\beta_B + j)B(1 + j, \beta_B)B(\alpha_A, \beta_A)}. \tag{6.57}$$

6.2 FREQUENTIST STATISTICS

The Frequentist paradigm (Definition 6.1) trace back to seminal works such as those of Neyman and Pearson [37] and Fisher [38], who laid the groundwork for much of its methodology. Subsequent developments by Wald [39], Neyman [40], and Lehmann [41] further refined its theories and techniques.
In the Frequentist paradigm, it is assumed that Nature's decisions can be captured by a statistical model with fixed, unknown parameters and an objective probability measure. In this setting, the optimal decision rule can be expressed as

$$U^*(x, w). \tag{6.58}$$

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 parameters 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 $x_{n+1} \in \Omega_X$ therefore takes the form[4]

$$U^*(x_{n+1}, \hat{w}^*(D)), \tag{6.59}$$

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 Definition 4.3 and Remark 6.2.

4 Equation 6.59 justifies the unified notation $U(\tilde{D})$ for the decision rule in Definition 4.3 and Remark 6.2.

6.2.1 *Frequentist Regression*

Regression involves the Robot constructing a model,

$$f : \Omega_\Theta \times \Omega_X \to \Omega_Y, \tag{6.60}$$

with associated parameters $\theta \in \Omega_\Theta \subseteq \Omega_W$, that estimates Nature's actions $y_{n+1} \in \Omega_Y$ based on observed data $x_{n+1} \in \Omega_X = \mathbb{R}$. 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 4.5 will be considered to review the subject.

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

$$p(y_{n+1}|x_{n+1}, w) = \sqrt{\frac{\xi}{2\pi}} e^{-\frac{\xi}{2}(f(\theta, x_{n+1}) - y_{n+1})^2}, \tag{6.61}$$

where $w = \{\theta, \xi, \dots\}$ denotes the collection of fixed, unknown parameters. Under the quadratic cost function from Definition 4.5, the optimal decision rule is the conditional expectation of Y_{n+1} given (x_{n+1}, w) (Theorem 4.3),

$$\begin{aligned} U^*(x_{n+1}, \hat{w}^*(D)) &= \mathbb{E}[Y_{n+1} \mid x_{n+1}, \hat{w}^*(D)] \\ &= \int y_{n+1} p(y_{n+1} \mid x_{n+1}, \hat{w}^*(D)) \mathrm{d}y_{n+1} \\ &= f(\hat{\theta}^*(D), x_{n+1}). \end{aligned} \tag{6.62}$$

Equation 6.62 represents the Frequentist optimal decision rule, defined conditional on an estimate of the model parameters. This can be directly contrasted with Equation 6.16, which expresses the Bayesian optimal decision rule obtained by averaging over the posterior distribution of the model parameters (and latent variables) given the observed data. From Equation 6.62, it follows that within the Frequentist paradigm, regression becomes a problem of parameter estimation.

6.2.2 *Frequentist Classification*

The setup for Frequentist classification initially mirrors that of Bayesian classification (Section 6.1.2), but start diverging at the level of the cost function

$$C(U(x_{n+1}, w), y_{n+1}) = 1 - \delta_{U(x_{n+1}, w), y_{n+1}} + (\lambda - 1)\delta_{U(x_{n+1}, w), \text{reject}}. \tag{6.63}$$

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_{n+1} is (Theorem 4.1 and Equation 4.43)

$$\begin{aligned} U^*(x_{n+1}, \hat{w}^*(D)) &= \underset{U(\tilde{D})}{\operatorname{argmin}} \mathbb{E}[C(U(\tilde{D}), Y_{n+1}) \mid x_{n+1}, \hat{w}(D)] \\ &= \underset{U(\tilde{D})}{\operatorname{argmin}} \sum_{y_{n+1} \in \Omega_Y} C(U(\tilde{D}), y_{n+1}) p(y_{n+1} | x_{n+1}, \hat{w}(D)) \\ &= \underset{U(\tilde{D})}{\operatorname{argmin}} \Big(1 - f_{U(\tilde{D})}(\hat{w}^*(D), x_{n+1}) + (\lambda - 1)\delta_{u,\text{reject}}\Big). \end{aligned} \tag{6.64}$$

From Equation 6.64, it follows that within the Frequentist paradigm, classification becomes a problem of parameter estimation.

Remark 6.3 (Bayesian versus Frequentist Regression and Classification)**.** *Contrasting Equation 6.62 and Equation 6.16 for regression and Equation 6.64 and Equation 6.34 for classification, the mathematical difference between the two paradigms can be written as shown in Table 1.*

Table 1: Comparison between Frequentist and Bayesian prediction

Paradigm	**Predictive model**
Frequentist	$f(\hat{w}^*(D), x_{n+1})$
Bayesian	$\mathbb{E}[f(\theta, x_{n+1}) \mid D]$

6.2.3 *Frequentist Parameter Estimation*

As shown in Section 6.2.1 and Section 6.2.2, 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 6.4 (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 $p(D \mid w)$.*

Remark 6.4 (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 $D \sim p(D \mid w)$ 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 $p(w \mid D)$ 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 6.3.

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 $D = \{(x_i, y_i)\}_{i=1}^n$ be the dataset. A bootstrap sample D^b 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:*

$$\widehat{\mathrm{Var}}_{\mathrm{boot}}[\hat{w}] = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{w}(D^b) - \overline{\hat{w}}^* \right)^2, \tag{6.65}$$

  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 6.5 (Fisher Information)**.** *Take $D_y = \{y_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ and let $w \in \Omega_W$ be an unknown parameter of the model. Let $p(D_y | D_x, w, I)$*

denote the likelihood of observing Natures actions D_y given observed data D_x and w. The Fisher information is defined as

$$\begin{aligned}\mathcal{I}(w) &\equiv \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_y|D_x,w)\right)^2\middle| D_x, w\right] \\ &= \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_y|D_x,w)\middle| D_x, w\right].\end{aligned} \tag{6.66}$$

Proof. In general

$$\mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p\right)^2\right] = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p\right] + \left(\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right]\right)^2. \tag{6.67}$$

Now

$$\mathbb{E}\left[\frac{\partial}{\partial w}\ln p\right] = \int\left(\frac{\partial}{\partial w}\ln p\right)p\mathrm{d}D_y \tag{6.68}$$

$$= \int\frac{\partial}{\partial w}p\mathrm{d}D_y \tag{6.69}$$

$$= \frac{\partial}{\partial w}\int p\mathrm{d}D_y \tag{6.70}$$

$$= 0, \tag{6.71}$$

since $\int p(D_y \mid D_x, w)\mathrm{d}D_y = 1$. Therefore

$$\mathcal{I}(w) = \mathrm{Var}\left[\frac{\partial}{\partial w}\ln p(D_y \mid D_x, w)\middle| D_x, w\right]. \tag{6.72}$$

□

Theorem 6.1 (Fisher Information for Independent Observations)**.** *Take $D_y = \{y_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ and let $w \in \Omega_W$ be a parameter of the model. Assume the likelihood factorizes as*

$$p(D_y|D_x,w) = \prod_{i=1}^n p(y_i|x_i,w). \tag{6.73}$$

Then, the Fisher information of the full dataset is

$$\mathcal{I}(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(D_y|D_x,w)\right)^2\middle| D_x, w\right] = \sum_{i=1}^n \mathcal{I}_i(w), \tag{6.74}$$

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

$$\mathcal{I}_i(w) = \mathbb{E}\left[\left(\frac{\partial}{\partial w}\ln p(y_i|x_i,w)\right)^2\middle| x_i, w\right]. \tag{6.75}$$

Definition 6.6 (Maximum Likelihood Estimator (MLE) Decision Rule)**.** *Take $D_y = \{y_i\}_{i=1}^n$, $D_x = \{x_i\}_{i=1}^n$ 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_y given D_x*

$$\hat{w}_{\text{MLE}}(D) \equiv \underset{w\in\Omega_W}{\operatorname{argmax}}\, p(D_y|D_x, w). \tag{6.76}$$

Theorem 6.2 (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*

$$\sqrt{n}\left(\hat{w}_{\text{MLE}} - w\right) \xrightarrow{d} \text{Norm}\left(0, \mathcal{I}(w)^{-1}\right), \tag{6.77}$$

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 6.7 (Minimax Decision Rule)**.** *A decision rule $\hat{w}'$ is said to be minimax if it minimize the maximum expected cost, meaning (Equation 4.17)*

$$\begin{aligned}\hat{w}' &\equiv \inf_{\hat{w}} \sup_{w\in\Omega_W} \mathbb{E}[C(\hat{w}, w)|w] \\ &= \inf_{\hat{w}} \sup_{w\in\Omega_W} \int C(\hat{w}(D), w)p(D|w)\mathrm{d}D.\end{aligned} \tag{6.78}$$

Theorem 6.3 (Mean Squared Error (MSE))**.** *The expectation of the quadratic cost function (Definition 4.5) can be written*

$$\begin{aligned}\mathbb{E}[C(\hat{w}, w)|w] &= \mathbb{E}[(\hat{w} - w)^2|w] \\ &= \mathbb{E}[(\hat{w} - \mathbb{E}[\hat{w}])^2|w] + (w - \mathbb{E}[\hat{w}])^2 \\ &= \text{Var}[\hat{w}|w] + \text{Bias}[\hat{w}|w]^2\end{aligned} \tag{6.79}$$

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

$$\text{Bias}[\hat{w}|w] \equiv w - \mathbb{E}[\hat{w}]. \tag{6.80}$$

If $\mathbb{E}[C(\hat{w}, w)|w] \to 0$ *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 6.1 (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 4.5), meaning it approximately minimizes the maximum expected cost.*

Proof. From theorem Theorem 6.3

$$\mathbb{E}[(\hat{w} - w)^2|w] = \text{Var}[\hat{w}|w] + \text{Bias}[\hat{w}|w]^2. \tag{6.81}$$

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

$$\begin{aligned} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w] &\approx \text{Var}[\hat{w}_{\text{MLE}}|w] \\ &\approx \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \end{aligned} \tag{6.82}$$

where Theorem 6.2 was used for the second line. The Cramer-Rao lower bound [42] for variance states that

$$\text{Var}[\hat{w}|w] \geq \frac{\text{tr}[\mathcal{I}(w)^{-1}]}{n}, \tag{6.83}$$

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

$$\sup_{w \in \Omega_W} \mathbb{E}[(\hat{w}_{\text{MLE}} - w)^2|w] \approx \inf_{\hat{w}} \sup_{w \in \Omega_W} \mathbb{E}[(\hat{w} - w)^2|w], \tag{6.84}$$

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

Example 6.4.

The bias-variance decomposition (Theorem 6.3) 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 6.5.

Take $D_y = \{y_i\}_{i=1}^n$ with $Y_i \sim \mathrm{Ber}(w)$, and let $w \in [0,1]$ be the unknown parameter. Determine the quadratic cost of three different decision rules for estimating w: the arithmetic sample mean, the constant 0.5, and the first observation y_1.

- *Arithmetic mean:*

$$\hat{w}(D_y) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.85}$$

with

$$\begin{aligned}
\mathbb{E}[\hat{w}(D_y)|w] &= \int \hat{w}(D_y)p(D_y|w)\mathrm{d}D_y \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i|w] \\
&= w, \\
\mathrm{Var}[\hat{w}(D_y)|w,I] &= \int (\hat{w}(D_y) - \mathbb{E}[\hat{w}(D_y)|w])^2 p(D_y|w)\mathrm{d}D_y \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[Y_i|w] \\
&= \frac{w(1-w)}{n}, \\
\mathbb{E}[(\hat{w}(D_y) - w)^2|w,I] &= \mathrm{Var}[\hat{w}(D_y)|w] + (\mathbb{E}[\hat{w}(D_y)|w] - w)^2 \\
&= \frac{w(1-w)}{n}.
\end{aligned} \tag{6.86}$$

- *Constant estimate:*

$$\hat{w} = 0.5 \tag{6.87}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}|w] &= 0.5, \\ \mathrm{Var}[\hat{w}|w] &= 0, \\ \mathbb{E}[(\hat{w}-w)^2|w] &= (0.5-w)^2. \end{aligned} \tag{6.88}$$

- *First observation:*

$$\hat{w}(D_y) = y_1 \tag{6.89}$$

with

$$\begin{aligned} \mathbb{E}[\hat{w}(D_y)|w] &= \mathbb{E}[Y_1|w] \\ &= w, \\ \mathrm{Var}[\hat{w}(D_y)|w] &= \mathrm{Var}[Y_1|w] + (\mathbb{E}[\hat{w}(D_y)|w] - w)^2 \\ &= w(1-w), \\ \mathbb{E}[(\hat{w}(D_y)-w)^2|w] &= w(1-w). \end{aligned} \tag{6.90}$$

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 y_1 is independent of n, making it less favorable as the sample size increases.

Example 6.6.

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

In this case

$$\begin{aligned} p(D_y|D_x, w) &= p(D_y|w) \\ &= \prod_{i=1}^{n} w^{y_i}(1-w)^{1-y_i}. \end{aligned} \tag{6.91}$$

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

$$\begin{aligned} \underset{w}{\mathrm{argmax}}\, l(w) &= \underset{w}{\mathrm{argmax}}\, p(D_y|w) \\ &= \underset{w}{\mathrm{argmax}} \ln\left(\prod_{i=1}^{n} w^{y_i}(1-w)^{1-y_i}\right) \\ &= \underset{w}{\mathrm{argmax}}\left[\ln w \sum_{i=1}^{n} y_i + \ln(1-w)\sum_{i=1}^{n}(1-y_i)\right] \end{aligned} \tag{6.92}$$

Now

$$\frac{d}{dw}l(w) = \frac{\sum_{i=1}^{n} y_i}{w} - \frac{n - \sum_{i=1}^{n} y_i}{1 - w} \tag{6.93}$$

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

$$\hat{w}_{MLE}(D_y) = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{6.94}$$

Example 6.7.

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

In this case

$$\begin{aligned} p(D_y|D_x, w) &= p(D_y|w) \\ &= \prod_{i=1}^{n} we^{-wy_i}. \end{aligned} \tag{6.95}$$

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

$$\frac{d}{dw}l(w) = \frac{n}{w} - \sum_{i=1}^{n} y_i \tag{6.96}$$

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

$$\hat{w}_{MLE}(D_y) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} y_i}. \tag{6.97}$$

APPENDIX A

Hamiltonian Monte Carlo

This appendix is taken from Petersen [43]. The Hamiltonian Monte Carlo Algorithm (HMC algorithm) is a Markov Chain Monte Carlo (MCMC) algorithm used to evaluate integrals on the form

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

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 [31]

$$g(\theta') = \int p(\theta'|\theta)g(\theta)d\theta. \tag{A.2}$$

Equation A.2 is implied by the stronger condition of detailed balance, defined viz

$$p(\theta'|\theta)g(\theta) = p(\theta|\theta')g(\theta'). \tag{A.3}$$

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 [44, 45]. 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 [30] 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 [32] 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

$$g = e^{\frac{F-H}{k_B T}} \Rightarrow H = F - k_B T \ln[g], \tag{A.4}$$

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

$$H = -k_B T \left[\ln[g] - \sum_i \frac{p_i^2}{2m_i} \right] + const, \tag{A.5}$$

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 [32], however, the temperature can be used to manipulate the range of states which can be accepted e.g. via simulated annealing [47]. Here $T = k_b^{-1}$ will be adopted in accordance with [31, 32] and as such

$$H = \sum_i \frac{p_i^2}{2m_i} - \ln[g]. \tag{A.6}$$

The dynamics in parameter space are determined by Hamiltons equations

$$\dot{\theta}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial \theta_i}, \tag{A.7}$$

with θ_i denoting the different variables (coefficients). In order to implement Hamiltons equations, they are discretized via the leap frog method [31, 32] viz

$$\begin{aligned} p_i\left(t + \frac{\epsilon}{2}\right) &= p_i(t) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t), p_i(t))}{\partial \theta_i}, \\ \theta_i(t+\epsilon) &= \theta_i(t) + \frac{\epsilon}{m_i} p_i\left(t + \frac{\epsilon}{2}\right), \\ p_i(t+\epsilon) &= p_i\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2} \frac{\partial H(\theta_i(t+\frac{\epsilon}{2}), p_i(t+\frac{\epsilon}{2}))}{\partial \theta_i}, \end{aligned} \tag{A.8}$$

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

$$\mathbb{P}(\text{transition}) = \min\left[1, e^{H_{initial} - H_{final}}\right]. \tag{A.9}$$

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 $\mathcal{O}(\epsilon^2)$ [31] 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

$$H = \frac{q^2}{2\sigma^2} + \frac{p^2}{2}. \tag{A.10}$$

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

$$\begin{bmatrix} q(t+\epsilon) \\ p(t+\epsilon) \end{bmatrix} = \begin{bmatrix} 1 - \frac{\epsilon^2}{2\sigma^2} & \epsilon \\ \epsilon(\frac{1}{4}\epsilon^2\sigma^{-4} - \sigma^{-2}) & 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \end{bmatrix} \begin{bmatrix} q(t) \\ p(t) \end{bmatrix} \tag{A.11}$$

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

$$\text{Eigenvalues} = 1 - \frac{1}{2}\epsilon^2\sigma^{-2} \pm \epsilon\sigma^{-1}\sqrt{\frac{1}{4}\epsilon^2\sigma^{-2} - 1}. \tag{A.12}$$

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

$$\epsilon < 2\sigma. \tag{A.13}$$

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

$$\epsilon_i < 2\lambda_i^{-\frac{1}{2}}, \tag{A.14}$$

for each eigenvalue λ_i of the matrix

$$M_{ij} = (\Sigma^{-1})_{ij}\frac{\partial^2 H}{\partial q_i \partial q_j}, \tag{A.15}$$

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

$$\epsilon_i < 2\sqrt{\frac{m_i}{\frac{\partial^2 H}{\partial q_i^2}}}. \tag{A.16}$$

Setting ϵ according to Equation A.14 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 [31, 32] (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] Edward E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, 1978, p. 25.

[6] David A. Freedman. *Statistics*. 4th. W. W. Norton & Company, 2007.

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

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

[9] Stephen Vineberg. "Dutch Book Arguments." In: *Stanford Encyclopedia of Philosophy* (2011). URL: https://plato.stanford.edu/entries/dutch-book/.

[10] Giuseppe Vitali. "Sui gruppi di punti e sulle funzioni di variabili reali." In: *Bologna: Zanichelli* (1905).

[11] D.H. Fremlin. *Measure Theory, Volume 1*. See Section 1E for the Vitali set construction. Torino, Italy: Torino: Torres Fremlin, 2000.

[12] Patrick Billingsley. *Probability and Measure*. 3rd. Chapter 1, Section on measure-theoretic foundations; Borel σ-algebra ensures well-defined probability measure. Wiley, 1995.

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

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

[15] E. T. Jaynes. “Probability Theory - The Logic of Science.”

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

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

[28] 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.

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

[30] 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.

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

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

[33] 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/.

[34] 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.

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

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

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

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

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

[44] 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.

[45] 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.

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

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

Index