INDUCTIVE REASONING UNDER CONSISTENCY

Marco C. Campi

printed May 26, 2024

Contents

Preface **3**

1 MODELS AND DECISIONS: CONCEPTS BY EXAMPLES **7**

1.1 Observation-Driven Models and Decisions 7

1.2 An overview of this monograph 10

2 PROBABILITY, INDEPENDENCE AND EXCHANGEABLE OBSERVATIONS **13**

2.1 Probability 13

2.2 Independent and exchangeable lists of observations 17

3 INTERPRETATION OF PROBABILITY **21**

3.1 Subjective probability 21

3.2 What is the role of subjective probability? 31

4 GUARANTEES AND PERSPECTIVES IN AN AGNOSTIC SETUP **35**

4.1 Observations as a means to create knowledge 35

4.2 Conditional knowledge 40

5 DECISIONS UNDER OBSERVATIONAL CONSISTENCY **45**

5.1 Models of a population 45

5.2 Decisions with consistency requirements 49

6 COMPLEXITY AND JUDGEMENTS **57**

6.1 Agnostic upper bounds to the risk 57

6.2 Assessments under non-degeneracy 66

6.3 A posteriori assessments 80

6.4 Drawing the conclusions: justifications and boundaries in the process of learning 81

The Epilogue **85**

Preface

Inductive reasoning refers to the process of synthesizing general principles, and interpretative models, from observations. This is key not only to the physical sciences, it also plays a fundamental role in large areas of applied fields such as engineering, medicine and economics. Even more broadly, induction forms the footing of all processes by which we learn from experience, with paramount implications in describing the world we live in and making decisions on how to operate on it. But, in virtue of what can observations be used to derive principles and to construct models meant to be applied to a new, out-of-sample, case? Why can observations drive our way of thinking and acting in situations that have not been previously encountered?

In this monograph, we pose and address these questions in a *formal*, *mathematical*, *language*. This comes with two advantages:

(i) mathematics demands to be precise about the hypotheses that we make and to be clear and explicit in the formulation of the conclusions. This is key to avoid the subtle pitfall of missing to clarify details and circumstances whose consideration would have shed a different light on the results or even shown internal inconsistencies in them;

(ii) the mathematical language is perfectly structured for making inference. This means that, by analytical methods, one can drive a long way to work out far-reaching results, which would otherwise be difficult to obtain.

It has to be said that any mathematical theory proceeds from premises to conclusions. Indeed, a formal language requires to first introduce the object the theory is intended to analyze and this process charges the object with specific properties and tells it apart from other entities that lie outside the scope of the theory. For example, in *group theory* one starts off considering a "group", that is, a set with an internal binary operation that obeys a given body of rules by which any two elements of the set are mapped into another element of the set. Integers $0, \pm 1, \pm 2, \dots$ with the sum operation is an example of group. Then, one proves theorems that are specific to the introduced setup. The theory of induction is no exception to this rule: one presents modeling premises meant to formalize inductive reasoning and investigates the implications they lead to.[1] Interestingly, the acceptance of the premises – and, thereby, the rational commitment to its implications – is on a voluntary basis, one accepts them inasmuch as they adhere to the way one reasons, or, in other cases, just to speculate on what a certain way of reasoning implies.

Inductive reasoning generates conclusions that acquire strength, or lose it, depending on the observations one is exposed to. Along the process, the conclusions are not certain, they have a given *degree of belief*. A proper tool to quantify the degree of belief is *probability theory* and this monograph is deeply grounded in probabilistic concepts. However, we do not require the reader to have any specific background in this discipline: the probabilistic tools necessary to comprehend the content of this work are introduced, and analyzed in their meaning, within this monograph. What this monograph does not contain are instead the proofs of the theoretical results. For this, the reader is referred, case by case, to the scientific publication in which the result has been demonstrated.

Some specific facts that are explored in this monograph are:

(a) in some contexts, probabilistic statements that certify the reliability of inductive conclusions can be formulated without any knowledge on the probability distribution by which observations are generated (*distribution-free* results). This corresponds to an *agnostic* point of view, where one admits the existence of a probability distribution but abstains from describing it. These findings shed light on the possibility of creating knowledge from experience, as opposed to, e.g., generating conclusions by blending probabilistic priors with observations, as is done in Bayesian inference;

(b) the time at which one speaks is crucial. Indeed, it makes a universe of difference making a reliability claim *before* the observations are actually collected and used (so that the claim refers to the inductive method, meant as the algorithm that maps observations into models) and *after* the application of the inductive method to a given set of observations (so that the claim refers to the outcome for the observations at hand). Both stances ("before" and "after") have practical interest and a clear separation between them help clarify intrinsic limits inherent in inductive reasoning.[2]

(c) when various individuals who bear different views get exposed to the same observations, their opinions tend to converge, with perhaps the exception of individuals who hold extreme views, incompatible with each others.

[1] This is not different from any other philosophical discourse even posed outside the mathematical language; however, at times, philosophical presentations conflate the process of deriving logical consequences with setting the stage for the premises from which these consequences follow.

[2] In a fortunate description due to Simone Garatti, these two stances have been related to the point of view of the *seller* and that of the *buyer*: the seller is interested in certifying the quality of the algorithm, so that he can set prices and policies, while the buyer is interested in the quality of the single item he is about to buy.

Throughout our exposition, we shall look for results that hold rigorously for any finite number of observations. This matters as one always uses in practice finite data sets; still, this is in contrast with much of the statistical literature, which is instead grounded on asymptotics, results that become valid only when the data set grows unbounded.[3]

Finally, some notes on the origins of this work and its limits of scope.

The idea of writing this monograph came to its author after some twenty years of mathematical investigation in the field of inductive reasoning. These studies have generated results that are believed to be of general interest to epistemologists, and yet the technical journals in which they have been published do not provide space for an in-depth presentation of the ensuing philosophical implications. Through this monograph, our intention is to fill this gap and position these findings so as to elucidate the importance we believe they have for a philosophical audience.

A comment is also due to explain the title of the monograph, "Inductive reasoning under consistency". The specification "under consistency" clarifies that this monograph is not omni-comprehensive, it pertains to specific reasoning procedures that involve the concept of *consistency*. Broadly interpreted, consistency refers to the property that a decision gets confirmed when it is appropriate for new incoming observations while it is refuted, and has to be changed in favor of a new decision, when confronted with incoming observations for which the initial decision is not appropriate. This framework applies broadly to models that incorporate a principle of parsimony of representation, as well as to large areas of observation-driven decision-making. However, it does not cover everything, chiefly it leaves out *least squares* approaches where the decision is determined through an averaging process (so that the decision changes whatever new observations are collected). Limiting the scope of this monograph this way is not a choice; more simply, at the time of writing this work, no theory is available that covers other frameworks to a similar depth. Nonetheless, it is essential to emphasize that the focus of this work is not comprehensiveness, rather it aims to highlight facts and principles by which inductive methods find a justification and to clarify intrinsic limits in the process of creating knowledge from observations.

A last thought goes to my co-author of all the technical facts exposed in this monograph. I wish to express my deep gratitude to Simone Garatti, his profound intelligence, generosity and dedication have been fundamental assets in a twenty-year voyage of enjoyable exploration in this field.

Milano, May 2024, Marco C. Campi

[3] This does not want to be an overall negative judgment of asymptotic results, which, by continuity arguments, may have a word in clarifying the effectiveness of specific methods on large data sets. Still, we maintain that results having rigorous validity for any finite number of observations offer a more robust foundation for addressing inductive problems.

Chapter 1

MODELS AND DECISIONS: CONCEPTS BY EXAMPLES

This first chapter aims to provide, by way of simple examples, a first acquaintance with various concepts relating to the generalization theory that will play a prominent role in this monograph. We first consider the problem of constructing a model to describe a population and, then, move to the more general setup of decision-making.

1.1 Observation-Driven Models and Decisions

§1 Observation-driven models. A first, primary, goal of inductive reasoning is that of building models meant to describe a whole population by the inspection of a restricted sample of members taken from the population. Here is an example.

EXAMPLE 1 (height and weight of Italians) *Suppose we are interested in the height and weight of Italians, so that Italians is our reference population and the height and weight of Italians are the two attributes we want to describe. Given that the height and the weight of a sample of Italians have been measured, a new Italian can be predicted to be not shorter than the shortest in the sample and not taller than the tallest, and, likewise, not lighter than the lightest or heavier than the heaviest in the sample.*[4] *This leads to constructing a model for the height and weight of Italians given by the rectangle* $[\min\{height_i\}, \max\{height_i\}] \times [\min\{weight_i\}, \max\{weight_i\}]$, *where index* $i = 1, \ldots, N$ *runs over the sample, and one next Italian is predicted to have height and weight so that it is represented by a point within the boundaries of this rectangle. See Figure 1.1 for an example.* $\ast$

[4] The use of the verb "can be predicted" indicates that this is a choice, and other possibilities would exist; for example, one might be willing to discard individuals whose somatic attributes are deemed extreme and unlikely to be met in the future.

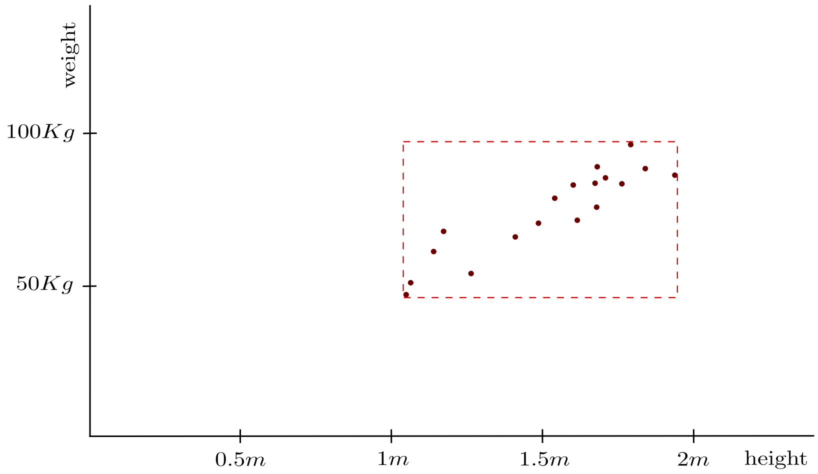

Figure 1.1: A rectangle describing the height and the weight of Italians. A new Italian is predicted to belong to the rectangle.

More generally, when speaking of *model of a population*, the world "population" does not necessarily refer to living beings, it has to be given a broad meaning. For example, it may refer to a collection of objects, or even to the set of conditions in which a machinery operates.

The usefulness of a model, like the one in the previous example, hinges on

(i) characteristics of the model itself (for example, the size of the spread of the interval used to predict an attribute); and

(ii) properties that relate to the interplay of the model with the population (what is the proportion of the members of the population that lie in the model?)

Point (ii) is relevant for judging the level of reliability of the model and may have significant consequences on our willingness to use it in applied problems. A notable difference between (i) and (ii) is the following: the model's characteristics can be directly inspected after the model has been built, whereas the reliability of the model remains hidden because it depends on the distribution of the population.[5] This raises fundamental *generalization questions*: is it possible to draw conclusions on members of the population we have not seen? how trustworthy are these judgments? what are the principles that logically support such a generalization process? These questions indeed point to the very essence of inductive reasoning.

§2 Decisions. More generally, inductive reasoning can be employed to *make decisions* with wide-ranging consequences for our way of acting and operating in the real world.

[5] If we knew the distribution of the population, the problem of finding a descriptor of the population from observations would disappear altogether.

EXAMPLE 2 (an investment problem) *Suppose that q assets $A^1,\ldots,A^q$ are available for trading. On period i, the asset A^j may gain or lose value in the market. Denoting p_i^j the closing price of asset A^j on period i, the ratio $\omega_i^j = (p_i^j - p_{i-1}^j)/p_{i-1}^j$ is called the* rate-of-return *of asset A^j on period i. It represents the percentage change in the value of asset A^j during period i. To manage uncertainty, investors diversify their portfolio. Thus, the investor will invest fractions $\theta^1,\ldots,\theta^q$ of his capital on $A^1,\ldots,A^q$ (we assume that $\theta^j \geq 0$ for all j, and $\sum_{j=1}^q \theta^j = 1$).*[6] *The vector $\theta = (\theta^1,\ldots,\theta^q)$ is called a "portfolio", and $\sum_{j=1}^q \theta^j \omega_i^j$ is the rate-of-return of the portfolio on period i: the investor increases/decreases his capital on period i by an amount equal to the rate-of-return of the portfolio per unit of money invested. The* portfolio loss *is simply the opposite of this quantity:*

$$L(\theta,\omega_i) = -\sum_{j=1}^{q} \theta^j \omega_i^j, \tag{1.1}$$

where ω_i is the vector $(\omega_i^1,\ldots,\omega_i^q)$.

Suppose now that the investor has actually observed n values $\omega_1 \ldots, \omega_n$ on previous trading periods. Substituting these values in (1.1)*, a set of n linear functions is obtained; see Figure 1.2 for a graphical representation in the case $q = 2$ and $n = 4$. This is called an "empirical bundle" of loss functions and it represents the observational wealth in the hands of the investor, who can employ it to make decisions on how to invest.*

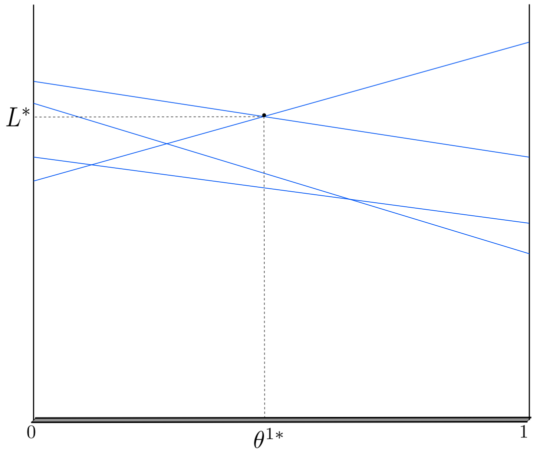

Figure 1.2: Graphical representation of the functions in (1.1). The abscissa represents θ^1 while θ^2 is obtained from relation $\theta^2 = 1-\theta^1$. Hence, the value $\theta^1 = 0$ corresponds to investing all capital on A^2 and $\theta^1 = 1$ to investing all capital on A^1.

As an illustrative example, consider min-max *optimization:*

$$\min_{\theta}\max_{i} L(\theta,\omega_i), \tag{1.2}$$

[6] Symbol $\sum$ stands for "sum". Therefore, $\sum_{j=1}^q \theta^j$ is a shorthand for $\theta^1+\theta^2+\cdots+\theta^q$.

which returns the portfolio selection θ^{1*} and the corresponding value L^* in the case of Figure 1.2.[7] *Suppose that the value L^* turns out to be satisfactory to the investor. Should he then invest according to θ^{1*}? Notice that real trading consists in first investing (i.e., choosing the value for θ^1 and θ^2) and only after the investment is made the market evolves, setting a value for ω that determines the final loss/reward. In contrast, the value L^* has been obtained by first observing $\omega_1 \ldots, \omega_n$ and then selecting θ^{1*} with the goal of optimizing the result for the cases that had been seen. This simply reverses the order one operates in reality. As a consequence, the value of L^* can be over-optimistic, and the obvious question is: how large is the chance that the actually incurred loss in a future investment will be larger than L^*? This is the problem of linking the "visible" to the "invisible", the past to the future, a generalization issue that is central to all inductive decision processes.* $*$

In this monograph, we investigate the grounds on which generalization in inductive processes can be rationally justified. Concepts and ideas will be expressed in the mathematical language whenever possible. While we understand that this is not customary to most of the philosophical literature, we believe it better serves the requirements of precision and clarity to which we feel obliged. In addition, the mathematical language – which is perfectly structured to make logical inference – will give us a helpful hand to run a tight ship while venturing in territories that can hardly be explored by other, less structured, means. On our side, we take full responsibility to introduce the mathematical tools, and elucidate their meaning, with care and gradualness, making them accessible even to the reader who only has a limited background in mathematics. Upon embarking on this journey, our hope is that we shall be able to convey to the reader at least half of the wonder we experienced when we first encountered the results we are about to narrate here.

1.2 An overview of this monograph

§3 Description by the chapters. The next two chapters are dedicated to the probabilistic concepts that are in use throughout this monograph. After introducing the mathematical definition of probability, Chapter 2 presents the concepts of *independent* and *exchangeable* observations, which play a central role in the temporal description of events. In turn, Chapter 3 carefully puts forward our interpretation of probability. While Chapter 3 is not technical, its importance should not be underestimated: it sheds the correct light under which all this monograph must be read. In Chapter 4, the reader makes a first encounter with some theoretical issues that are central to inductive methods. By way of simple examples, this chapter guides through an exploration of the

[7] *Min-max* corresponds to a conservative attitude that places all emphasis on the worst (the investor selects the portfolio that minimizes the loss on the worst trading periods). This approach is seldom applied to real trading. We use it here because it allows us to illustrate some concepts more easily; see Example 10 in §36 for a more realistic approach.

intrinsic boundaries within which inductive reasoning proceeds, coming to the conclusion that agnostic reasoning (in which probabilistic beliefs are built from observations without any prior) is possible. Chapter 4 prepares the reader for the following two chapters, which form the theoretical core of this work: Chapter 5 introduces a broad framework for inductive modeling and decision-making centered around the concept of *consistency*, and the following Chapter 6 presents the ensuing generalization theory. More specifically, after formally introducing the concept of "consistent rule", Chapter 5 provides various examples to better appreciate its meaning and applicability. In turn, Chapter 6 furnishes an ample account of the generalization results that are applicable to a consistent framework. By introducing an essential separation between an inductive statement and the reasoning by which this statement is obtained, this chapter justifies the process of learning from observations while also identifying the intrinsic limits this process has necessarily to meet.

Chapter 2

PROBABILITY, INDEPENDENCE AND EXCHANGEABLE OBSERVATIONS

Inductive reasoning is deeply grounded in probability theory, which provides the tools used to describe uncertain knowledge. This chapter introduces the probabilistic concepts that are relevant to this monograph. The presentation is self-contained and easily accessible without any specific background knowledge.

2.1 Probability

§4 Elementary probability. Consider a set Ω that contains a finite number of elements. In this monograph, we always regard Ω as the set of the possible outcomes of an experiment (*observations*).[8] For example, when modeling coin tossing, we may want to choose $\Omega = \{head, tail\}$ and, when throwing a die, $\Omega = \{1,2,3,4,5,6\}$. To each element $\omega \in \Omega$, we associate a real number $\mathbb{P}(\omega) \in [0,1]$, called the *probability* of ω, in such a way that the sum of all such numbers adds up to the value 1: $\sum_{\omega\in\Omega} \mathbb{P}(\omega) = 1$. Any subset E of Ω is called an *event*. The probability is extended to events by the definition $\mathbb{P}(E) = \sum_{\omega\in E} \mathbb{P}(\omega)$. This defines a map, called *probability function* or *probability distribution*, that associates to any subset of Ω a number in $[0,1]$, its probability. The probability function is clearly *additive*, in the sense that $\mathbb{P}(E_1 \cup \cdots \cup E_m) = \mathbb{P}(E_1) + \cdots + \mathbb{P}(E_m)$,[9] whenever $E_1, \ldots, E_m$ are *incompatible*, i.e., they are disjoint sets.

[8] More generally, in probability theory Ω may also contain hidden variables that are not directly accessible through experiments.

[9] Symbol $\cup$ stands for "union". Therefore, $E_1 \cup \cdots \cup E_m$ indicates the set of $\omega \in \Omega$ that belong to at least one of the sets E_1, E_2, …, E_m.

§5 A general definition of probability. While elementary probability is easy to understand and therefore holds pedagogical value (indeed, we shall resort to elementary probability to explain various concepts in many parts of this monograph), it is unable to cover needs that emerge in multiple contexts. More generally, set Ω may contain infinitely many elements, and one may like to give the status of event to a reduced collection of subsets of Ω, rather than all subsets of Ω. These generalizations are captured by the axiomatic definition of probability introduced in 1933 by Andrej N. Kolmogorov, [46], as explained in the following.

Given an arbitrary set Ω, let us consider a collection of subsets $\mathscr{E}$ of Ω that has the following properties:

(i) $\Omega \in \mathscr{E}$ (the whole set Ω, called the *universe*, is in $\mathscr{E}$);

(ii) if $E \in \mathscr{E}$, then $E^c \in \mathscr{E}$ (if a set E is in $\mathscr{E}$, then its complement E^c, formed by all elements of Ω but those in E, is also in $\mathscr{E}$);

(iii) if $E_1, E_2 \in \mathscr{E}$, then $E_1 \cup E_2 \in \mathscr{E}$ (if E_1 and E_2 are in $\mathscr{E}$, then their union is also in $\mathscr{E}$).

A collection of sets that satisfy (i)-(iii) is called an *algebra* and $\mathscr{E}$ is referred to as the *algebra of events*. It is easy to show that (i)-(iii) imply that also the union and the intersection of any finite collection of sets in $\mathscr{E}$ is in $\mathscr{E}$.

Further, to each $E \in \mathscr{E}$, one associates a real number, called its *probability* and indicated with the symbol $\mathbb{P}(E)$, in such a way that the following properties are satisfied:

(iv) $\mathbb{P}(E) \in [0,1]$ for any $E \in \mathscr{E}$, and $\mathbb{P}(\Omega) = 1$ (the probability of the universe is 1);

(v) given any finite or infinite list of disjoint events $E_1, E_2, \dots$ such that their union $E = \cup_i E_i$ is also in $\mathscr{E}$,[10] it holds that $\mathbb{P}(E) = \sum_i \mathbb{P}(E_i)$ (this property is known as *σ-additivity*).

As can be easily verified, elementary probability as in §4 is just a simple instance of the general framework introduced here.

Properties (i)-(v) define Kolmogorov's axiomatic system of probability. Using the same terminology as in the elementary case, the map $\mathbb{P}$ that associates a number in $[0,1]$ to any element of $\mathscr{E}$ is called the *probability function* or the *probability distribution*. Given Ω and $\mathscr{E}$, there are clearly many possible choices of probability functions that satisfy Kolmogorov's axiomatic system. For example, if $\Omega = \{1,2,3,\dots\}$ (the set of natural numbers) and $\mathscr{E}$ contains all subsets of Ω, then a possible choice is to give probability $\frac{1}{10}$ to the first 10 natural numbers and zero to all others (which, by property (v), defines the probability of all other events); another choice that attributes non-zero probability to all natural numbers is that of giving probability $\frac{1}{2}$ to the first natural 1,

[10] This needs to be specified because the infinite union of events need not necessarily be an event, i.e., it can be that $\cup_i E_i \notin \mathscr{E}$.

probability $\frac{1}{4}$ to the second natural 2, probability $\frac{1}{8}$ to the third natural 3, and so on, or, in more compact writing, $\mathbb{P}(z) = \frac{1}{2^z}$ to any $z \in \Omega$.

It has been said that Kolmogorov's axiomatic system contains elements that are unessential – and even undesirable – in a definition of probability because these elements are not motivated by any practical reason when probability is used to model real situations. Firstly, one may wonder why the collection of events $\mathscr{E}$ to which a probability is associated must be an algebra, rather than just a generic collection of sets. Secondly, σ-additivity applies to an infinite sequence of events, which hardly reflects any realistic requirement in practical usage since infinity is not found in the real world.[11] It is a fact that σ-additivity limits the applicability of the theory because, otherwise well-conceived, candidate probability functions have to be discarded on the ground that they do not satisfy the σ-additivity property. On the other hand, using the axiomatic apparatus proposed by Kolmogorov provides the undeniable advantage that it casts probability within the well-established wake of measure theory (in which σ-additivity plays a central role), which grandly simplifies its use. This is indeed the main reason of the great success of Kolmogorov's approach. Throughout this monograph, we shall make exclusively use of Kolmogorov's axiomatic system of probability. In §15, we shall come back to this choice when commenting on de Finetti's definition of probability, a framework that is more general and strictly contains that of Kolmogorov.

Beyond mathematical definitions, it is important to highlight that the concept of probability has many interpretations, as witnessed by a truly vast and multiform scientific and philosophical literature. We make clear at this early stage that, throughout this monograph, probability will be given only one, well-identified, meaning as *subjective probability*, as we shall discuss in full detail in the next chapter.

§6 Elementary conditioning. Consider two events A and C in $\mathscr{E}$ with $\mathbb{P}(C) \neq 0$. The *conditional probability* of A given C is defined as $\mathbb{P}(A|C) = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)}$.[12] The idea is that one takes the probability of A under the condition that C also happens, which gives $\mathbb{P}(A \cap C)$, and then normalizes such a probability by dividing by $\mathbb{P}(C)$. It is easy to see that, for any given C, $\mathbb{P}(A|C)$, seen as a function of $A \in \mathscr{E}$, is itself a probability distribution. Often, the operation of conditioning is used to describe how knowledge updates: what is the probability of A if I know that C has happened? This interpretation is discussed in §16. More generally, one may want to condition on events C that have zero probability. Probability theory admits such an operation, and this topic is dealt with in

[11] See, e.g., the thought-provoking article [17]; in there, we read: "*If we, on the basis of a convention, state that $\mathbb{P}(E_i) = 0$, $i \geq 1$, entails $\mathbb{P}(\cup_i E_i) = 0$, then we intuitively think of $\cup_i E_i$ as a nearly impossible event, whereas the formal definition allows us only to conclude that 0 is the value at $\cup_i E_i$ of the function which we, conventionally, have called probability.*" Kolmogorov himself seems to be aware of this arbitrariness; in [46], he writes: "*Since the new axiom [σ-additivity] is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning [$\cdots$]. We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI.*"

[12] Symbol $\cap$ stands for "intersection". Therefore, $A \cap C$ indicates the set of $\omega \in \Omega$ that are simultaneously in A and C.

any textbook on probability. Interestingly, when considering zero-probability events, the mathematical notion of conditioning may introduce extra elements that cannot be directly traced back to the interpretation that one initially associates to the elementary operation of conditioning (particularly, due to the fact that conditional probability can be redefined arbitrarily on events of probability zero).[13] We feel fortunate that these issues do not affect to any significant degree any of the concepts we deal with in this monograph.

§7 Elementary image probability. Consider the set $\Omega = \{1,2,3,4,5,6\}$ as in §4 with the probability distribution given by $\mathbb{P}(\omega) = \frac{1}{6}$ for any $\omega \in \Omega$. Suppose we place a bet on even numbers, 2, 4 and 6, so that we receive one unit of money if ω is even and lose one unit of money if ω is odd. This defines a function f from Ω to the set $\{-1,1\}$ that maps $1,3,5$ in -1 and $2,4,6$ in 1. Its inverse f^{-1} is a multi-valued map that returns the subset of values in Ω that are brought by f to the chosen target value, either -1 or 1. For example, $f^{-1}(1) = \{2,4,6\}$. Function f^{-1} allows us to introduce a probability distribution (let us indicate it with the symbol $\mathbb{P}'$) on the subsets of $\{-1,1\}$ defined by the following two relations: $\mathbb{P}'(-1) = \mathbb{P}(f^{-1}(-1))$ and $\mathbb{P}'(1) = \mathbb{P}(f^{-1}(1))$. Clearly, this gives $\mathbb{P}'(-1) = \mathbb{P}'(1) = \frac{1}{2}$. $\mathbb{P}'$ is called the *image probability distribution* and, in our example, it may be interpreted as the probability of losing/winning one unit of money, see §16. The notion of image probability distribution can be carried over to the context of functions defined on a generic set Ω as in §5, with some care for so-called *measurability issues*[14] (the interested reader can consult any textbook on probability for a full treatment).

§8 Elementary composition probability distribution. Suppose that a set contains only two elements, α and β. We probabilize the subsets of $\{\alpha,\beta\}$ according to two alternative possibilities as follows: $\mathbb{P}_1(\alpha) = 0.4$; $\mathbb{P}_1(\beta) = 0.6$ or $\mathbb{P}_2(\alpha) = 0.7$; $\mathbb{P}_2(\beta) = 0.3$. Further, we also assign a probability distribution to the subsets of $\{1,2\}$, say $\pi(1) = \pi(2) = 0.5$. Using these ingredients, we can construct a probability distribution on the domain of the pairs formed by an element taken from $\{1,2\}$ and a second element taken from $\{\alpha,\beta\}$: $\mathbb{P}(i,\alpha) = \mathbb{P}_i(\alpha)\cdot\pi(i)$; $\mathbb{P}(i,\beta) = \mathbb{P}_i(\beta)\cdot\pi(i)$, $i = 1,2$. For example, $\mathbb{P}(1,\alpha) = 0.4\cdot 0.5 = 0.2$, this is the probability of α in condition 1 times the probability of condition 1 and it is often interpreted as the probability of the simultaneous happening of 1 and α, see §16. This concept can be extended to arbitrary sets using the notion of *Markov kernel*, with due attention to measurability issues, as discussed in any good textbook on probability theory.

[13] Even more intriguing, according to definitions of probability alternative to the one due to Kolmogorov (we are referring in particular to non σ-additive probabilities, see, e.g., [56]), conditioning to zero-probability events may produce values that are systematically strictly smaller than the value obtained by directly conditioning on the union of all these events even when this union has a non-zero probability, a manifestation of so-called *non-conglomerability* (see [3] for a comprehensive treatment of this topic). The fact that advanced mathematical definitions raise interpretative doubts is in no way a specificity of probability theory; for example, in the theory of partial differential equations the concept of derivative in weak solutions may as well give rise to interpretative issues.

[14] These functions are called *random variables.*

2.2 Independent and exchangeable lists of observations

§9 Independent and identically distributed (i.i.d.) lists in the elementary case. We first consider the setup of elementary probability in which Ω contains a finite number of elements.

Consider lists $(\omega_1, \omega_2, \ldots, \omega_n)$ of n elements ω_i from Ω with repetition (i.e., the same element can appear more than once).[15] We want to introduce a probability function on the space of all such lists. One (easy) way is to define the probability of a single list as $\mathbb{Q}(\omega_1, \omega_2, \ldots, \omega_n) = \Pi_{i=1}^{n}\mathbb{P}(\omega_i)$.[16] Thus, the probability of a list is just the product of the probabilities associated with each element in the list. As it can be easily verified, the sum of the probability of all lists equals one. Then, the probability of any set of lists is defined by summing up the probabilities of all lists in the set. This defines a probability on lists in full analogy with the definition of probability on subsets of Ω as done in §4.

We can now state some simple properties of $\mathbb{Q}$. Introduce the notation $(E_1, E_2, \ldots, E_n)$ to represent all lists $(\omega_1, \omega_2, \ldots, \omega_n)$ in which $\omega_1 \in E_1, \omega_2 \in E_2, \ldots, \omega_n \in E_n$. Then, one has $\mathbb{Q}(E_1, E_2, \ldots, E_n) = \Pi_{i=1}^{n}\mathbb{P}(E_i)$.[17] (The fact that probability $\mathbb{Q}$ can be broken up as a product probability is referred to as that each component in the list is *independent* of the others.[18]) It immediately follows that $\mathbb{Q}(\Omega, \ldots, \Omega, E_i, \Omega, \ldots, \Omega) = \mathbb{Q}(\Omega, \ldots, \Omega, E_j, \Omega, \ldots, \Omega)$, whenever E_i and E_j are the same set but placed in a different position in the list[19] (this is expressed by saying that the components are *identically distributed*).

§ 10 Independent and identically distributed lists. The treatment for general probabilities follows the same path as for elementary probability, with just a bit of attention for technical details.

In the elementary case, single lists played the role of fundamental building blocks, and matters of convenience suggested to start from them. Instead of single lists, in the general case one starts off by considering $(E_1, E_2, \ldots, E_n)$, the set of all lists whose first component is in E_1, whose second component is in E_2, and so on, where each E_i is a set in $\mathscr{E}$.[20] One defines $\mathbb{Q}(E_1, E_2, \ldots, E_n) = \Pi_{i=1}^{n}\mathbb{P}(E_i)$. However, the collection

[15] In inductive reasoning, a list is interpreted as an ordered collection of observations.

[16] Symbol Π stands for "product". Therefore, $\Pi_{i=1}^{n}\mathbb{P}(\omega_i)$ is a shorthand for $\mathbb{P}(\omega_1)\mathbb{P}(\omega_2)\cdots\mathbb{P}(\omega_n)$.

[17] For example, take $n = 2$ and let $E_1 = \{\omega^1, \omega^2\}$ and $E_2 = \{\omega^2, \omega^3\}$, where $\omega^1, \omega^2, \omega^3$ are three elements of Ω. Then, $\mathbb{Q}(E_1, E_2) = \sum_{k=1,2}\sum_{j=2,3}\mathbb{Q}(\omega^k, \omega^j) = \sum_{k=1,2}\sum_{j=2,3}[\mathbb{P}(\omega^k)\cdot\mathbb{P}(\omega^j)] = [\sum_{k=1,2}\mathbb{P}(\omega^k)]\cdot[\sum_{j=2,3}\mathbb{P}(\omega^j)] = \mathbb{P}(E_1)\cdot\mathbb{P}(E_2)$.

[18] The terminology "independent" finds its motivation in the interpretation of conditional probability as knowledge updating, see §16. To understand this, introduce the notation $\{\omega_1 \in E_1\}$ to indicate all lists in which the first element is in E_1, that is, $\{\omega_1 \in E_1\} = (E_1, \Omega, \ldots, \Omega)$. If, e.g., $n = 2$ and $\mathbb{P}(E_1) \neq 0$, then $\mathbb{Q}(\{\omega_2 \in E_2\}|\{\omega_1 \in E_1\}) = \frac{\mathbb{Q}(\{\omega_1 \in E_1\}\cap\{\omega_2 \in E_2\})}{\mathbb{Q}(\{\omega_1 \in E_1\})} = \frac{\mathbb{Q}(E_1, E_2)}{\mathbb{P}(E_1)} = \frac{\mathbb{P}(E_1)\cdot\mathbb{P}(E_2)}{\mathbb{P}(E_1)} = \mathbb{P}(E_2)$, which is interpreted that observing $\omega_1 \in E_1$ does not alter the probability that $\omega_2 \in E_2$.

[19] Indeed, the two sides equal $\mathbb{P}(E_i)$ and $\mathbb{P}(E_j)$ respectively, which are equal because $E_i = E_j$.

[20] Starting from single lists is inappropriate for various reasons, in particular a singleton $\{\omega^i\}$ need

of all sets $(E_1, E_2, \ldots, E_n)$ that are obtained as the E_i's vary in $\mathscr{E}$ does not form an algebra. To obtain an algebra, one considers all finite unions of such disjoint sets: $\cup_{i=1}^{m}(E_{1,i}, E_{2,i}, \ldots, E_{n,i})$, where m is any positive integer and all sets $E_{j,i}$ are chosen arbitrarily from $\mathscr{E}$ in such a way that $(E_{1,i}, E_{2,i}, \ldots, E_{n,i})$, $i = 1, \ldots, m$, are disjoint.[21] To each of these unions, one attributes probability $\sum_{i=1}^{m} \mathbb{Q}(E_{1,i}, E_{2,i}, \ldots, E_{n,i})$ and, by a cumbersome calculation, one can show that this $\mathbb{Q}$ is σ-additive.[22]

§11 Exchangeability. Exchangeability is more general than "independence with identical distribution", which we discussed in §10, and contains it as a particular case.

Over the algebra in the space of lists that has been described in §10 (given by all finite unions of disjoint sets of the form $(E_1, E_2, \ldots, E_n)$), consider any probability distribution $\mathbb{Q}$ such that $\mathbb{Q}(E_1, E_2, \ldots, E_n) = \mathbb{Q}(E_{i_1}, E_{i_2}, \ldots, E_{i_n})$ holds for any permutation $(i_1, i_2, \ldots, i_n)$ of the indexes $(1, 2, \ldots, n)$.[23] Then, the lists are said to be *exchangeable*.

Clearly, i.i.d. lists are exchangeable because in the i.i.d. case $\mathbb{Q}(E_1, E_2, \ldots, E_n)$ is computed as a product, which renders the order of the factors inessential. The converse is only partially true: exchangeable lists are identically distributed, however they are not independent in general. To see that they are identically distributed, one takes $\mathbb{Q}(E, \Omega, \ldots, \Omega)$. As E varies in $\mathscr{E}$, this gives the probability distribution of the first component. By the property of invariance under permutation, one obtains, for example, that $\mathbb{Q}(E, \Omega, \ldots, \Omega) = \mathbb{Q}(\Omega, E, \Omega, \ldots, \Omega)$, so that the probability distribution of the first component equals that of the second. Similarly, one shows that all components share the same probability distribution. The fact that exchangeability is more general than independence is shown by means of an example.

EXAMPLE 3 (Pólya's urn) *A Pólya's urn, named after the Hungarian mathematician George Pólya, is an urn that contains balls colored in two ways: red and white. One ball is drawn from the urn and its color is observed. It is then returned in the urn, and an additional ball of the same color is added to the urn. We suppose that the urn contains initially two balls, one red and one white, and that we repeat the drawing process n times. While we do not venture here to explain the exact interpretation we attribute to probability in this experiment (for, we have postponed describing our interpretation of probability altogether until the next chapter), we take Pólya's urn just as an intuitive expedient to introduce the following probability distribution over lists. Given a list, say, $(r, w, r, r, \ldots, w)$ ("r" stands for red and "w" for white), its probability is given the value $\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{4} \cdot \frac{3}{5} \cdot \ldots \cdot \frac{*}{n+1}$ because in the first draw there are in the urn two balls one of which is red, in the second draw there are three balls and only one is white, in the third draw there are four balls and two are red, and so on (the*

not be in $\mathscr{E}$ and, hence, there might be no probability associated to it.

[21] Showing that the collection of all unions $\cup_{i=1}^{m}(E_{1,i}, E_{2,i}, \ldots, E_{n,i})$, as m runs over positive integers and $E_{j,i} \in \mathscr{E}$, is an algebra is a simple and instructive exercise.

[22] In the following, to indicate the probability distribution $\mathbb{Q}$ we shall often use the symbol $\mathbb{P}^n$, which helpfully recalls the fact that $\mathbb{Q}$ is a "product probability distribution".

[23] A permutation of $(1, 2, 3)$ is $(2, 1, 3)$, another permutation is $(3, 1, 2)$.

reader will figure out what "$*$" stands for). Clearly, components are not independent since previous draws affect the probability of subsequent draws. However, they are exchangeable. Indeed, by a re-arrangement of terms, one easily see that the probability of a list in which the numerosity of reds is $\#(r)$ and that of whites is $\#(w)$ is given by $\frac{\#(r)!\cdot\#(w)!}{(n+1)!}$,[24] which does not depend on the order in which red and white balls are drawn. $*$

Arguably, the property that the probability distribution of the components does not change through time has connections with David Hume's "Principle of Uniformity", [41, 40]. We shall come back to this point in §17.

By and large, this monograph is centered around the concept of independence, rather than exchangeability. The main reason for this choice is that the theory of inductive reasoning under consistency has been amply developed for independent lists of observations, while the corresponding theory for exchangeable lists has not grown mature at the time this monograph is being written. Nonetheless, to the author of this work this limitation does not seem to have any severe implication because the main messages this monograph is meant to convey can well be cast within the independent framework.

[24] Given an integer q, the symbol $q!$ (to be read "q factorial") indicates the number $q\cdot(q-1)\cdots 2$.

Chapter 3

INTERPRETATION OF PROBABILITY

In science and philosophy, the concept of probability has been used in diverse contexts and a multitude of alternative interpretations have been associated to it. Consequently, when employing probabilistic concepts, it is essential to explicitly define what probability represents. This chapter is entirely dedicated to explaining how probability must be interpreted in this monograph, and it plays a key role in ensuring a proper understanding of the entire work.

3.1 Subjective probability

§12 The necessity of declaring the interpretation of probability. Probability is ubiquitous. It plays an essential role in engineering, underlies much of the social sciences, figures prominently in quantum and statistical mechanics, and takes center stage in financial and actuarial studies. Nonetheless, how probability has to be interpreted in a given context is often not obvious and claiming that it is an "idealization" is simplistic and it is no excuse to using it without providing an explicit statement about what it is meant to represent. To make this point clear beyond any doubt, we feel advisable to illustrate it by means of a simple example. To model the water level in a tank with an orifice at its bottom, I can employ the ordinary differential equation

$$\frac{\mathrm{d}\ell(t)}{\mathrm{d}t} = -\sqrt{\ell(t)} + q(t),$$

where $\ell(t)$ is a function of time that represents the water level, symbol $\frac{\mathrm{d}}{\mathrm{d}t}$ denotes the operation of derivation that describes how a quantity varies with time (for example, velocity of a car is the derivative of its position), $q(t)$ is the water in-flow rate and $-\sqrt{\ell(t)}$ describes the out-flow rate through the orifice according to Torricelli's law. If, e.g., we take $\ell(0) = 1$ (tank initially full at level 1) and $q(t) = 0$ for any t (no water inflow), the differential equation yields the solution $\ell(t) = 1 - t + \frac{1}{4}t^2$ for $t \leq 2$ and $\ell(t) = 0$ afterward, a function that starts from value 1, initially decreases fast, then slows down and touches zero (empty tank) at time $t = 2$. I don't believe that this is a perfect description of the evolution of the water in the tank, and the differential equation captures only an idealized version of the phenomenon. Still, it is clear what it is meant to describe: the water level in the tank, a quantity that I can measure with a yardstick. This clarity of interpretation is not germane to probability: what do we exactly mean when we say that the probability of a certain event to happen tomorrow is 90%? or when we say that the probability of throwing an ace with a die is $1/6$? Is this our belief or rather something that is inherent in the object under consideration? Since the word "probability" can be used with different meanings, it is our duty to declare what probability is meant to describe when we use it.

§ 13 Subjective probability. Let us consider an individual (myself)[25] who has partial knowledge about a phenomenon. Suppose that the phenomenon can instantiate in two ways, A or B (for example, team A or team B will win tonight's match). I am unsure about which of the two will happen but, still, I feel that the chance of A is higher than that of B. This feeling is called *degree of belief*; how degrees of belief are built is discussed in §14. Probability can then be used to express my degree of belief: I attribute value $2/3$ to A and $1/3$ to B or, perhaps, if my degree of belief on A is even higher, $3/4$ to A and $1/4$ to B.

Importantly, the degree of belief describes an intimate interplay between me and the phenomenon, also in the light of the knowledge I have on it. For example, if someone has thrown two dice and I can't see any of them, I may give probability $1/36$ to the outcome $(1,1)$; however, if I see one die and it indeed shows an ace, then I may give probability $1/6$ to $(1,1)$. Since lack of knowledge itself justifies partial belief (*epistemic uncertainty*), the use of probability does not assume – nor does it exclude – intrinsic uncertainty that cannot be compensated for by experience (*aleatoric uncertainty*).

The above interpretation of probability, named *subjective probability*, has been proposed and developed independently by Bruno de Finetti, [21] and [23], and Frank P. Ramsey, [55], in the first half of the 20th century.[26] We shall give more space to de

[25] While I do not necessarily embrace a solipsistic vision, which would lead me into a territory for which I feel unprepared, I hold that there is no reason for me to assume the existence of others throughout my discussion on the interpretation of probability. Hence, probability, as described here and used throughout this monograph, refers to myself, to the way I think. Nevertheless, I shall often indulge in expression like "one" or "we" (not only as *pluralis majestatis*) to indicate bearers of probabilistic beliefs. This is an expedient that helps when it comes to comparing the consequences of various probabilistic beliefs I can have (and, moreover, indulging in referring to others makes me feel less lonely in the endeavor of writing this monograph, a weakness for which I beg the reader's indulgence).

[26] While de Finetti and Ramsey have championed subjective probability, already almost a century before in 1847 Augustus De Morgan, [51], had written: "By degree of probability, we really mean, or ought to mean, degree of belief".

Finetti's approach in §15.

In this monograph, *we shall exclusively interpret probability according to the aforementioned subjective point of view*. While we need not to justify this choice, as it is our choice, we also make explicit that we do not adamantly exclude that probability can be satisfactorily employed with other interpretations.[27] But even so, what matters is that probability is an appropriate instrument to describe partial beliefs, and their evolution as new information is acquired, which is the subject matter of this monograph.

§ 14 How degrees of belief are built. Degrees of belief are created in various ways. (i) I can analyze an object to form an opinion on how it operates. For example, if I analyze a coin and deem it to be a true coin with no manipulations, then I can assign equal probability of 0.5 to it landing heads or tails; (ii) previous trials can suggest probabilistic values. Suppose I execute an experiment 1000 times and record the empirical frequency with which the outcome takes one of three possible values a, b, c. I can then use these relative frequencies as the probability to obtain an a or b or c at the next trial (more on the relation between frequencies and probabilities in §18); (iii) probabilistic values can be suggested by a person I interact with, who claims to have an experience on the phenomenon at hand; (iv) in the presence of outcomes that exhibit apparent symmetry (for example, receiving a given set of five cards in a card game), one can appeal to a *principle of indifference* (a terminology coined by John M. Keynes) and assign the same probability to each outcome[28]; (v) yet another approach, proposed by Bruno de Finetti, consists of relating probabilities to odds in a wager: the probability of an event is a fair price to enter a bet on that event. We shall give a close attention to this approach in the next §15.

§ 15 Obedience to the laws of probability. The axioms of probability provide a *mathematical model* for the degrees of belief. I am not obliged to abide by these axioms, I can or cannot accept them depending on whether my way of thinking aligns with them and, if I accept them, they provide the foundation to construct a theoretical framework for describing the evolution of beliefs by using *deductive logic*. This is not different from accepting or not accepting a frictionless model for describing the movement of a cart, from which certain logical conclusions can be drawn. Hence, *the axioms of probability need not be justified or proven*; rather, they should be chosen in a way that guarantees widespread acceptance because, otherwise, they lose interest.

[27] We do not reject, for instance, that probability can be used to describe an intrinsic indetermination in physical quantities beyond any possibility to compensate for it by observations, as is held in quantum mechanics. Henri Poincaré would disagree with us, in [52] he writes: "*If we were not ignorant, there would be no probabilities; there would only be room for certainty*". However, delving into a thoroughgoing presentation of the interpretations of probability alternative to the subjective one simply lies outside the scope of this monograph.

[28] While inchoate versions of this principle were already present in Blaise Pascal, Jacob Bernoulli and Gottfried W. von Leibniz, the principle of indifference was fully developed into a theoretical apparatus mainly by Abraham de Moivre, [24], and, later, by Pierre S. Laplace, [48].

Kolmogorv's definition of probability in §5 has become mainstream in probabilistic studies. Most of its prescriptions are broadly acceptable according to a subjective point of view: particularly, it seems reasonable that the probability of the universe is 1 and that the probability with which one among two mutually exclusive events happens is the sum of the probabilities of the two events. However, the requirements that the set of events $\mathscr{E}$ form an algebra and the property of σ-additivity are more questionable. It is a fact that alternative formalizations of the concept of probability have been proposed that eschew these requirements, and we feel obliged to briefly describe the one proposed by Bruno de Finetti as, we believe, it adds value to our discussion.[29]

De Finetti introduced a framework for *coherent bets*, that is, bets that do not entail a sure loss, and proved that the axioms of probability can be logically derived within this framework. Consequently, in this perspective, the axioms of probability emerge as a consequence of coherence, rather than serving as the starting point of the theory. To be specific, consider any collection $\mathscr{E}$ (finite or infinite) of subsets from a universe Ω. The members of $\mathscr{E}$ are called *events*. For any event $E \in \mathscr{E}$, let us assign a price $p(E)$ for entering a bet where one receives one unit of money if E proves true and zero otherwise. The opponent can purchase bets on multiple, albeit finitely many, events $E_1, \ldots, E_q$ and pay us the corresponding prices, while he can simultaneously force us to buy additional bets on $E_{q+1}, \ldots, E_{q+m}$ under analogous, but reversed, conditions. In other words, we set the prices, but the opponent decides which side will be ours in each bet. For example, if the opponent purchases the bet on E_1 and gets us to buy the bet on E_2, we immediately receive the sum $p(E_1) - p(E_2)$ (either positive or negative). Subsequently, we pay one unit of money to the opponent if E_1 proves true while E_2 does not; conversely, we receive one unit of money from the opponent if E_2 proves true while E_1 does not. No payment occurs in the other cases (both E_1 and E_2 are true or both are not true). A set of prices is deemed *coherent* if it prevents the opponent from constructing sets of multiple bets in which we certainly incur a loss. This implies that the prices are additive: if $E_1, E_2 \in \mathscr{E}$ are incompatible and $E_1 \cup E_2$ is also in $\mathscr{E}$, then $p(E_1 \cup E_2) = p(E_1) + p(E_2)$. To see this, suppose that $p(E_1 \cup E_2) < p(E_1) + p(E_2)$ (the opposite case is similar). If the opponent bets on $E_1 \cup E_2$ and gets us to bet on E_1 and E_2, he immediately receives the net positive amount $p(E_1) + p(E_2) - p(E_1 \cup E_2)$, and certainly there is no exchange of money afterward, so that this situation contradicts coherence. Coherence also implies that the price for betting on the universe is 1. What coherence does no imply is that the set $\mathscr{E}$ is an algebra, nor does it imply the property of σ-additivity. Next, de Finetti argues that the prices we assign to bets mirror our personal degrees of belief. Indeed, if I believe that event E is likely to occur, I will demand a high price to bet on it; however, I cannot exaggerate, as doing so would entail an expected loss should the opponent force me to bet on it. Therefore, probabilities are equated with prices.[30]

[29] De Finetti first proposed his ideas in [21], while a classic, extended, exposition is [22].

[30] Identifying probabilities with prices may be viewed as a stretch, and indeed it has raised perplexities. For instance, Edwin T. Jaynes, [43] writes "*it seems to us inelegant to base the principles of logic*

As it may have appeared, we lean towards sympathizing with de Finetti's probability framework because it avoids an (artificial) use of the mathematical concept of *infinite* (as it does the property of σ-additivity). On the other hand, it is a fact that the whole theory of inductive reasoning under consistency, which is the central focus of this treatise, has been developed in technical papers within the tradition of Kolmogorov's probability, and we don't dare reposition it here outside this established framework. Therefore, *throughout we shall make reference to the axiomatic system of probability described in §5*. While asking for the reader's mercy for this choice, we also notice that the σ-additivity does not affect in its essence any of the fundamental ideas that we mean to put forward in this work and, therefore, we suggest taking σ-additivity just as an idealization at the service of mathematical simplicity.

We conclude this point by noting that the axioms of probability delimit the feasible domain within which a probability distribution can be chosen, but the actual selection of probabilistic values rests with the individual, who makes a choice guided by background knowledge, also in the light of personal inclinations.[31] In a fortunate simile, de Finetti himself compared a probability distribution to a drawing, wherein the laws of perspective correspond to the axioms in probability theory; in [23], he writes: "*If someone draws a house in perfect accordance with the laws of perspective but choosing the most unnatural point of view, can I say that he is wrong? [$\cdots$] the various internally consistent opinions about probabilities can analogously be conceived as all the perspectives obtainable by varying our point of view*".

§ 16 Conditioning and image/composition probability. The operation of conditioning has been described in §6. In subjective probability, conditioning has often been advocated as a mathematical tool that describes the updating of beliefs when a new piece of information comes along. For example, suppose that I believe that having two consecutive days of sun, tomorrow and the day after, has probability 45% while the probability of just having one sunny day, tomorrow, is 60% and so is the probability that it is sunny the day after. So, before any observation, I hold the belief that in two days from now I'll have a sunny day with probability 60%; however, if I wait until tomorrow and tomorrow is a sunny day, then I update my belief to the probability $\mathbb{P}(\{\text{day after tomorrow sunny}\}|\{\text{tomorrow sunny}\}) = \frac{\mathbb{P}(\{\text{day after tomorrow sunny}\}\cap\{\text{tomorrow sunny}\})}{\mathbb{P}(\{\text{tomorrow sunny}\})} = \frac{45\%}{60\%} = 75\%$.[32]

Similarly to §15, I am not compelled to adopt conditioning as a rule for updating

on such a vulgar thing as expectation of a profit". For example, a Buddhist monk may have little interest in money, so much so that he can accept any set of prices, while he may carry strong beliefs on the world. We here take the defense of de Finetti for the same reason that an axiomatic probability system has to be accepted, not justified: de Finetti's model will be accepted, and put at work, only when one agrees to identify probabilities with prices as a reasonable modeling assumption.

[31] This point is further commented upon in §19, where we briefly discuss logical probability.

[32] Some authors contend that conditional probability should only be interpreted as a state of mind prior to actually seeing any observations: "*if I would see that ...*". We do not embrace this point of view, and take conditioning as a *factual* updating rule that can be used after actually collecting an observation.

probabilities: it is a model, and I will accept it provided I see it describes my way of thinking.[33]

Instead, the image probability distribution, as discussed in §7, is a tool to introduce a probability distribution on the co-domain of a function whose domain already admits a probability distribution. Domain and co-domain describe two separate entities, and – while sensible and easily acceptable – assuming that we build a probability distribution on the co-domain as an image probability distribution is again a modeling assumption that I can, or cannot, abide by. Similarly, the composition probability distribution discussed in §8 is a model of the degree of belief in the simultaneous occurrence of two events, and its acceptance is, once more, a modeling assumption.

Throughout this monograph, we adhere to the modeling interpretation of conditioning and image/composition probability distribution as described in this section. Consequently, for instance, we assume that the image probability distribution is an accepted model of our beliefs in induced phenomena, without having to explicitly declare it each time.

§17 Exchangeable observations. In §11 we noted that the notion of exchangeability has connections with David Hume's "Principle of Uniformity". We come back to this observation here and discuss it in some detail. In [40], Hume questioned whether one can rationally use observations collected in the past to predict the future. He argued that this requires that the future will resemble the past, i.e., that the course of nature is *uniform*. However, he denied the possibility of rationally drawing the conclusion that nature is uniform with his famous two-pronged criticism: (a) the uniformity of nature cannot be proven deductively because "*it implies no contradiction that the course of nature may change, and that an object, seemingly like those which we have experienced, may be attended with different or contrary effects*"; (b) it cannot be proven by referring to experience either since "*all our experimental conclusions proceed upon the supposition that the future will be conformable to the past*", hence using experience would lead to circularity. Therefore, according to Hume, no reasoning can justify forming conclusions that go beyond the past instances of which we have had experience. This dilemma is famously known as "the problem of induction".

As we have seen in §11, exchangeability implies the invariance of the probability distribution (and so does the i.i.d. assumption of §10), which can be interpreted as an attempt to formalize the principle of uniformity.[34] It is of the utmost importance,

[33] When walking down a long path in probabilistic computations, it is not credible that we describe how we think; rather, we are describing how we would think, should we be endowed with enough mental power. Additionally, the exercise of deriving conclusions from premises by mathematical manipulations that use accepted probabilistic rules suggests me what I must think if I want to be consistent with the adopted probabilistic model.

[34] In a probabilistic framework, the probability distribution is a complete description that assigns a probability to each single event. Let us consider, for example, a simple set Ω that only comprises two elements: $\{$radionuclide R decays within a time T from its generation$\}$ and its opposite $\{$radionuclide R does not decay within a time T from its generation$\}$. Assuming invariance of the prob-

however, that exchangability is a property of the probability distribution of lists of observations, and the only meaning given in this monograph to probability is subjective. Therefore, *assuming exchangability in no way posits a state of nature, it merely puts forward a modeling assumption on the way I expect that the flow of observations unfolds*, an assumption that I can or cannot accept depending on the circumstances as dictated by the problem at hand. *Our primary interest in this work is to investigate – by the only use of deductive logic – how knowledge advances when a model that assumes exchangeable observations is accepted.* This endeavor will bring us into a territory of deep exploration, e.g., about the possibility of creating knowledge from observations in a condition of initial absence of probabilistic prejudgments, or about the convergence of the opinions of (hypothetical) individuals who hold distinct initial ideas and get exposed to common experiences.

Before concluding this point, we find it advisable to add that accepting the invariance of probabilities is not a condition we have, or have not, to accept universally; in fact, its acceptance is *highly dependent on the specific problem that we are dealing with*. It has been argued, as seen in [27], that laws with temporal restrictions would be inherently more mysterious and puzzling than ones that are temporally universal. While this can be true for fundamental laws of physics, it is important to recall that inductive reasoning is an essential tool in applied fields spanning from economics to engineering, from medicine to meteorology, from control to social sciences. In all these disciplines, one looks at specific portions of the real world, and I want to argue that adopting such a partial perspective can introduce an apparent time-variability, which we need to recognize in our process of modeling. To understand this point, consider the following system of differential equations:

$$\begin{cases} \frac{\mathrm{d}x_1(t)}{\mathrm{d}t} & = & i_1(t) \\ \frac{\mathrm{d}x_2(t)}{\mathrm{d}t} & = & x_1(t) \cdot x_2(t) + i_2(t), \end{cases}$$

where $i_1(t)$ and $i_2(t)$ are two inputs. This is clearly a time-invariant system in the sense that it reacts to external stimuli and initial conditions independently of when we start operating on it. In fact, suppose that the system is initialized at time $t = 0$ with $(\bar{x}_1, \bar{x}_2)$ and is fed with the inputs $\bar{i}_1(t)$ and $\bar{i}_2(t)$, and let $(\bar{x}_1(t), \bar{x}_2(t))$ be the corresponding *movement* (i.e., $(\bar{x}_1(t), \bar{x}_2(t))$ is the solution to the system of differential equations). If we now postpone the initialization until time $\tilde{t}$: $(x_1(\tilde{t}), x_2(\tilde{t})) = (\bar{x}_1, \bar{x}_2)$ and apply the delayed version of the inputs $\bar{i}_1(t - \tilde{t})$ and $\bar{i}_2(t - \tilde{t})$, then the corresponding movement becomes $(\bar{x}_1(t - \tilde{t}), \bar{x}_2(t - \tilde{t}))$, that is, the same movement as before with the only difference that it is delayed by $\tilde{t}$ instants, which we interpret as time-invariance. On the other hand, suppose that, say, $i_1(t) = 1$ for any t and that $\bar{x}_1 = 0$. By integrating the first equation $\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = i_1(t)$, we obtain $x_1(t) = t$. Substituting this solution in the

ability distribution means that the probability of decaying within time T remains constant over time, allowing us to conceive of learning this probability through experiments conducted in a laboratory.

second equation now gives

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = t \cdot x_2(t) + i_2(t),$$

in which time t appears explicitly; hence, in the partial perspective of the second equation, the system looks now time-varying.[35] This phenomenon of apparent time-variability is ubiquitous in applied sciences and may challenge our willingness to accept a model that postulates the invariance of the probability distributions. For example, when dealing with the rates-of-return of financial assets, we may assume uniformity of behavior over a limited time window, while it is common opinion that the behavior of the market does change across longer periods of time. Still, it has to be noted that entire fields in telecommunications, machine leaning, control and information theory, to cite but a few examples, regularly assume the invariance of the probability distributions. Therefore, studying inductive reasoning within this framework is well worth it not only in connection with fundamental laws but also to deal with problems in more applied fields.

§ 18 Repeated experiments. I instinctively prefer certainty to uncertainty and, hence, concentrated probability distributions are more palatable because they are closer to certainty. Interestingly, concentration in the distribution often emerges out of repetition. For example, suppose that a string of symbols has probability 4% of getting corrupted in the transmission through a channel. Correspondingly, the universe contains the following two elements: $a = \{\text{the string gets corrupted}\}$ and $b = \{\text{the string does not get corrupted}\}$, where the first has probability 4% and the second 96%. To describe repeated, independent experiments, we use the procedure described in §10. Hence, the probability of a list $(\omega_1, \omega_2, \ldots, \omega_n)$, where each ω_i is either a or b, is given by $(0.04 \cdot \#(\omega_i = a)) \cdot (0.96 \cdot \#(\omega_i = b))$. Suppose, for example, that $n = 10000$ and consider $f = \frac{\#(\omega_i = a)}{10\,000}$, the empirical frequency of corrupted strings. A calculation that involves Bernoullian distributions shows that the distribution of f is concentrated around 0.04 so much that it belongs to the interval $(0.032, 0.048)$ with probability 99.99%.[36]

Importantly, the above result must be given exclusively a subjectivist interpretation: if my degree of belief in a is 4%, then my degree of belief that f is in $(0.032, 0, 048)$ is 99.99%. This result has nothing to do with the frequentist interpretation of probability along an approach first proposed by John Venn, [66], and then developed, among others, by Richard von Mises, [67], and Hans Reichenbach, [57], according to which the probability of an event is the limiting relative frequency with

[35] For example, if the equation is initialized at time $t = 0$ with $\bar{x}_2 = 1$ and is fed with the input $\bar{i}_2(t) = 0$ for any t, then $\bar{x}_2(t) = \exp(\frac{1}{2}t^2)$, te exponential of $\frac{1}{2}t^2$. However, delaying the initialization until time $\tilde{t} = 1$ ($x_2(1) = 1$) and applying the delayed version of the input $\bar{i}_2(t-1) = 0$ for any $t \geq 1$ yields $\tilde{x}_2(t) = \exp(\frac{1}{2}t^2 - \frac{1}{2})$ for $t \geq 1$; this is different from the delayed version of the movement $\bar{x}_2(t)$, that is, $\bar{x}_2(t-1) = \exp(\frac{1}{2}(t-1)^2)$.

[36] This is the image probability of f according to §16.

which such an event occurs in repeated trials. While this latter interpretation leads to a number of inconsistencies that have already taken up too much time of many gifted researchers, we just want to dismiss it here as being uninteresting (because of its infinitary nature) and poorly defined (because the mutual relation among trials are not, and cannot be, precisely characterized).

§ 19 Logical probability. While we mentioned earlier that a thoroughgoing exploration of the various interpretations of probability goes beyond the scope of this monograph, we feel advisable to digress momentarily on the notion of *logical probability* because this can better position our use of subjective probability.

Early proponents of logical probability were John M. Keynes, [45], and Harold Jeffreys, [44], but the most unyielding supporter of this interpretation has been Rudolf Carnap, [15]. Logical probability holds that any piece of evidence confers an *objective* support (or *confirmation*) to given hypotheses. The relation between evidence and hypothesis is logical, and probability extends sure, deductive logic to uncertain, inductive logic. Quoting directly from [15]: "*Deductive logic may be regarded as the theory of the relation of logical consequence, and inductive knowledge as the theory of another concept which is likewise objective and logical, viz, probability*$_1$ *[Carnap calls "probability*$_1$*" the logical probability, in contrast to the frequentist probability, which he calls "probability*$_2$*"] or degree of confirmation. That probability*$_1$ *is an objective concept means this: if a certain probability*$_1$ *value holds for a certain hypothesis with respect to a certain evidence, then this value is entirely independent of what any person may happen to think about these sentences, just as the relation of logical consequence is independent in this respect.*" He adds: "*Suppose somebody makes the statement in deductive logic: 'h follows logically from j.' [...] The statement 'the probability of h on the evidence e is* $\frac{1}{5}$*' has the general character as the former statement; therefore it cannot violate empiricism any more than the first. Both statements express a purely logical relation between two sentences.*"

Let me now express my personal take on logical probability. Logical probability can perhaps be viewed as a *model of perfect thinking* in the absence of certainty. If so, in a sense logical probability is akin to our subjective probability, for we have said that subjective probability is a model of my thinking. However, the problem with this interpretation of logical probability is that it is not clear what it is meant to be a model of, that is, what "perfect thinking' means. Is it perhaps a mysterious entity whose existence we are here positing? Altogether, it seems to me that associating perfect thinking to something beyond the circularity of what logical probability itself defines turns out to be difficult. Alternatively, logical probability can be seen as a *rational way of thinking*, which extends deductive logic, and indeed this is the interpretation put forward by Carnap. However, if I am dragged into discussing this point of view, which I reluctantly do, I must say that it is not obvious to me that even deductive logic can be given a "logical status". Take the sentence: if all elements in a set A have property P, then all elements of a set B that is contained in A have property P (a relation between two propositions established by deductive logic); isn't this a transposition of the experience that if we remove a bucket of balls from a box that contains all red balls, then, upon inspection, all balls in the bucket are red? While I do not want to open a discussion on deductive logic here, certainly I cannot accept that probability theory is a logical extension of deductive logic that extends, and is justified, beyond experience, for not even deductive logic has this status in my mind. On the other hand, deductive logic is quite specific and so well delimited that I feel authorized to "freeze" it and treat it as being given beyond any reasonable doubt as part of my model for inductive reasoning, while probability theory refers to such a multitude of situations and conditions that coming to a consolidated ground is impossible, and indeed this is the very reason why probability theory only introduces general rules that limit the way a probability distribution can be chosen.

Beyond all the foregoing conceptual criticisms, what might be the practical effect of assuming logical probability in a scientific theory of induction?[37] In whatever conception of logical probability, determining the actual values of the degree of confirmation for the claims that stand in a certain relation to evidence is practically impossible. Then, lacking a quantitative support, the role of logical probability in a scientific theory of induction, however conceived, becomes doubtful. But there is more than that. Assuming that my probabilistic judgments somehow follow – or even, to a certain degree, conform to – a perfect way of judging may encumber my ability to freely speculate on the very reasons by which probabilistic reasoning is effective, which is detrimental to the development of a scientific theory of induction. Altogether, the role of logical probability seems to us more akin to that of God in a religion, its existence is reassuring and certainly has a value, but this value lies outside the domain of free, scientific, speculation.[38]

§ 20 Sets of probabilities. Probability is a suitable instrument to describe my thinking in conditions of uncertainty. On the other hand, *there is no reason why we should only use probability, warding off the possibility of mixed descriptions, partly probabilistic and partly set-theoretic*. For example, if I know that in an urn there are 100 balls, partly red and parly white, but I do not have any evidence of the mutual proportion of the two, why should I not model my belief by saying that the probability of extracting a red ball in the next draw is $\frac{m}{100}$, where m is any number between 0 and 100?[39]

[37] By "scientific theory of induction" we mean a free speculation that pursues logical consequences stemming from a prescribed model of inductive reasoning.

[38] Certainly, we do not deny the right of philosophy to speak of God and other metaphysical entities. We simply observe that introducing logical probability in a scientific theory of induction may hinder one's ability of free speculation.

[39] In Bayesian probability, all elements present in the problem are probabilized according to subjective beliefs. As a consequence, the subjective interpretation of probability is largely adopted in Bayesianism, which has spawned the widespread misconception that subjectivism implies the use of Bayesian probability.

While using a set of probabilities seems to us a perfectly legit way of modeling, at times the principle of indifference has been advocated to create fully probabilistic models. However, the principle of indifference may generate inconsistencies and should be applied with care. The following example taken from [65] illustrates the idea. A factory produces wooden squares with variable side-length between 0 and 1 meter. If I have no information about the production process, by advocating the principle of indifference I may assume uniform probability distribution of the length. Then, I conclude that the probability of drawing a square with side-length between 0 and 0.5 meters is 50%. Let us re-formulate our model. A factory produces wooden squares with variable area between 0 and 1 square meter. Since I have no information about the production process, I advocate the principle of indifference and assume uniform probability distribution of the area, leading to the probability 50% of drawing a square with area between 0 and 0.5 square meters. This result is in contradiction with the previous conclusion, which leads to probability 50% of drawing a square whose area is between 0 and 0.25 square meters (to draw this conclusion, use the image probability distribution for $area = length^2$ obtained from the probability distribution of the length, see §16). Clearly, this inconsistency is generated by there being more than one way to carve up the space of alternatives, which show that adopting the principle of indifference may imply choices that surreptitiously drive our evaluations.

Beyond the principle of indifference, it is a plain fact that in many problems we nowadays address in applied fields such as finance, medicine and engineering, assuming that we hold probabilistic beliefs on all uncertain elements present in the problem is completely unrealistic. For instance, when modeling the effect of a defibrillator in the resuscitation of an individual in cardiac arrest, I well accept that a probability distribution describes the physiological conditions of the individual and the ability of the defibrillator to resuscitate him. On the other hand, describing this probability distribution, or providing probabilistic weights for the various probability distributions that I can envisage as done in a Bayesian approach, does not appear to be a reasonable modeling methodology. Interestingly, theories can be conceived that are able to generate results of practical utility without introducing any knowledge on the possible distributions by which cases are generated (*distribution-free theories*). Hence, the existence of a probability distribution is a necessary mental condition to apply the theory, but the actual assignment of the probabilistic values is not required to use the ensuing results. The theory of inductive reasoning under consistency that we shall present later in this monograph is (in a sense to be carefully specified) one such theory.

3.2 What is the role of subjective probability?

§21 Knowledge and decision processes. If probabilistic beliefs are purely subjective, why are they so important? one might ask. The simple answer is that they describe my expectations on events when full, certain knowledge is not possible. They have, in a sense, the same character as certain knowledge insofar they describe my understanding and take on things, but also have the additional flexibility to account for the level of trust I hold when full trust is not reasonable. This knowledge matters as it is my knowledge, and it plays an essential role in those deliberations that are intended to guide my practical decisions.

Correspondingly, the impact of probabilistic beliefs is quite broad, and indeed it extends beyond what is commonly thought. It is normal experience that only partial and imprecise knowledge is available in many endeavors of everyday life. On the other hand, many fields in decision-making involving financial, medical and engineering evaluations are the realm of uncertainty, which is the sole reason for the quest of a trade-off between robustness and performance. But there is more, and indeed all physics is based on "consolidated probabilistic beliefs": physical laws are nothing but the crystallization of relations among natural variables on which our belief is so high that it is practically convenient to treat them as certain. They are not proven to be universally valid, they are just conventionally deemed to be universally valid, and they drive our way of thinking and acting.[40]

§22 What is then left to an engineer? What is then left to an engineer who is interested in a design that must work when applied to the world we perceive? We do not have an answer at this stage, simply because the perceived world is not in the picture yet. Probabilistic beliefs describe a state of mind, and probability theory models its evolution under the effect of observations. The observations are the inputs to the model, but how they are generated is in no way contained in the probabilistic model. On the other hand, it makes perfect sense to complement a probabilistic model of beliefs with a model of the world we perceive and ask how the two interact. The best explanation is by an example.

EXAMPLE 4 (RandN) *In the manual of* Matlab*, a software commonly used for mathematical computations, the routine* RandN *is described as a generator of normally distributed random numbers*[41] *with zero mean and unitary variance. And this is what I hold in my mind. Suppose that my interest lies in the average* $\frac{x_1+x_2+\cdots+x_{50}}{50}$ *of lists of* 50 *independent numbers* $(x_1,x_2,\ldots,x_{50})$ *generated with* RandN. *It is known that the average of Gaussian and independent random variables is also a Gaussian random variable. By a simple calculation, I conclude that its mean is zero and its variance equals* $1/50$ *(this is the image probability distribution of the average obtained from* $(x_1,x_2,\ldots,x_{50})$ *according to the procedure in §16). As a consequence, the average belongs, e.g., to the interval* $(-0.42,0.42)$ *with probability* 99.73%. *This is what*

[40] De Finetti, referring to Henri Poincaré, wrote in [23] "*he has clearly understood that only an accomplished fact is certain, that science cannot limit itself to theorizing about accomplished facts but must forsee, that science is not certain, and that what really makes it go is not logic but the probability calculus.*"

[41] A normally distributed random number is a number generated from a random variable that has a Gaussian distribution.

I hold in my mind, it is my belief. How does it relate to the actual average I obtain when I really generate 50 *numbers with* RandN*? This question finds no answer in my probabilistic model simply because the way* RandN *really operates is not in the picture yet. It is a fact that* RandN *operates on the ground of a* seed*, whose value can also be specified by the user and, once the seed is given, the list is generated by a fully deterministic procedure. I can then consider a large amount of seeds, say* 10 000*, look at the corresponding lists of* 50 *numbers, and verify for how many, and possibly which, seeds the ensuing average indeed falls in the interval* $(-0.42, 0.42)$. $*$

This example was about an extremely simple situation. However, it makes perfect sense to employ the same conceptual scheme in complex, real-life problems. Given any probabilistic method that proceeds using partial information to draw conclusions, one can wonder how this method will work in practice. This question cannot be answered until one introduces a model of the perceived world, describing how it generates observations and how it reacts to the stimuli we impose on it.[42,43] To accomplish this descriptive task, our preference[44] is to adopt a deterministic approach because, otherwise, we would find ourselves in the awkward position of having to clarify to which part of the perceived world the introduced probability refers. It also makes quite a bit of sense to assume multiple deterministic alternatives to describe the perceived world so as to safeguard against various occurrences and possible time-variability, and it can be reassuring to see that the adopted procedure works well in many of the envisaged alternatives.[45]

§23 About the existence of the real world. A voice suggests us to briefly clarify the implications of §22 on our perspective on the *real world*. Indeed, there aren't any. When modeling the perceived world, we are not assuming that the real world exists in any given form, we do not give any ontological value to our action of modeling. Describing the perceived world is simply a mental exercise to verify the quality and effectiveness of a probabilistic procedure in relation to our perception of the real. To me, it makes perfect sense to say: if I describe the operation of the world I perceive

[42] One can also operate empirically by a verification of the actual effectiveness of the conclusion. However, a factual verification cannot testify to the correctness and quality of the probabilistic method (in whatever sensible sense one can give to the words "correctness" and "quality"), it only puts in relation a specific action, or judgment, determined from a probabilistic method with factual observations.

[43] At times, these two modeling levels – what I think and the model of how the world operates – are made to coincide. For example, when an engineer designs a bridge, he may make reference to principles of structural engineering that are deemed certain in first approximation (see §21), which constitutes both his beliefs and the model of the world.

[44] We say "preference" because, as already noted at the end of §13, we have no fundamental reasons to exclude other possibilities.

[45] At times, when the perceived world is assumed to operate in one among various ways, we reduce our judgment about the adopted procedure to a single number by averaging (with weights adding up to one and describing, perhaps, the relative importance we attribute to cases) the performances obtained over the possible operating conditions. Despite the weights having the mathematical structure of a probability distribution, one should refrain from giving any probabilistic interpretation to this operation.

this way and I accrue knowledge on it and process it this other way, then the overall structure formed by the model of the world I perceive and my operating on it behaves according to the following scheme. It is just a dissociation of my thinking between how the world I perceive is described and how I come to learn about it and operate on it. *This position stands in the face of the obvious objection that I can only think my thoughts*. Making clear this view matters because this monograph does not want to provide any contribution to the debate on the existence and nature of the real world beyond our perceptions: this monograph takes a distance from any support, or refutation, and indeed any judgment on the nature of reality.

Chapter 4

GUARANTEES AND PERSPECTIVES IN AN AGNOSTIC SETUP

This chapter is not yet about inductive reasoning under consistency (which is the subject of the next two chapters), rather it is preparatory to it. Through the examination of simple examples, it is shown that precise reliability claims are possible in an agnostic setup in which no preliminary probabilistic knowledge is assumed. However, a fundamental distinction must be drawn between judging the reliability of an inductive procedure and evaluating the reliability of the procedure outcomes for a given sample of observations. This distinction is explored in the second section of this chapter under the title of "conditional knowledge".

4.1 Observations as a means to create knowledge

§24 The agnostic setup. Using the rules of probability, probabilistic beliefs can be updated into new probabilistic beliefs. At times, however, I start from a set of probabilities, rather than one single probability distribution, so as to accommodate my subjective inability to describe uncertainty by way of one single probabilistic stand (see §20). The extreme situation is encountered when any probability distribution is deemed possible, so embracing a *fully agnostic* attitude.[46] If so, is it still possible to draw conclusions of theoretical value – and practical interest – by using probabilistic methods? As mentioned, for the largest part of this monograph, we shall adopt

[46] While in this monograph the word "agnostic"simply means that *any probability distribution of the observations is possible*, we also note that our agnostic analysis well conforms to the ethos of Thomas H. Huxley; in [42], he writes: "*Agnosticism is not a creed, but a method, the essence of which lies in the rigorous application of a single principle [...]: in matters of intellect, follow your reason as far as it will take you, without regard to any other consideration.*"

the mental stance that observations follow an i.i.d. (independent and identically distributed) scheme, which reflects our idea that the problem at hand exhibits invariant properties over time. Within this framework, in subsequent chapters we shall present deep-seated and fully agnostic results in relation to the broad area of inductive reasoning under consistency. In the current chapter, we focus on a simpler framework that allows us to put forward some initial, and yet fundamental, facts about the possibility of making agnostic judgments.

The value of the exploration in an agnostic setup rests in the fact that *there is a substantial difference between adopting the mental stance that an uncertain phenomenon can be described by a probability distribution and assuming that we are able to assign the probabilistic values of this distribution (we already encountered this same idea in §20).* In fact, when grappling with problems involving complex and articulated data generation mechanisms, assuming the availability of a description of the underlying probability distribution is often unrealistic. At a deeper conceptual level, updating a distribution only describes how *a priori* knowledge morphs into updated beliefs as new observations are acquired. Nonetheless, this does not resolve the fundamental quandary of inductive reasoning: the origin of knowledge. In fact, *a priori* knowledge remains in need of justification even after explaining the process of updating beliefs based on new observations. In contrast, *generating informative probabilistic results in an agnostic setup goes deep into exploring the mechanisms by which knowledge can be created out of ignorance in the light of observations, and this points to the very core of inductive reasoning.*

§25 More about i.i.d. and agnosticism. While this point repeats ideas that have been already introduced elsewhere in this monograph, we feel it advisable to give them a unified presentation in the interest of clarity.

Our theoretical framework for inductive reasoning adopts a minimalist approach, involving only two elements: observations and beliefs. Observations drive the evolution of beliefs in a continuous flow, where observed facts are used to update beliefs and beliefs are employed to anticipate facts that have not yet been observed. How observations are generated is not part of our discourse; we refrain from positing any ontological status for the entity that generates observations. Therefore, a probability is always subjective and only used to describe beliefs. Particularly, adopting an agnostic stance corresponds to not restraining my subjective judgments about uncertainty in any possible way.

In this context, introducing an i.i.d. scheme for the flow of observations simply represents a mental attitude, not a postulation about the real world. It is not *descriptive*; instead it *prescribes* a subjective model of how facts are believed to unfold over time. One is free to accept or not to accept this model, and its consequences will be regarded as valuable to the extent that one honestly specifies it. Certainly, an inductive process cannot proceed without assuming some connections between past and future, seen and unseen. The hypothesis of i.i.d. serves this purpose, and alternatives, perhaps less demanding, are well possible. They are not explored in this monograph simply because of the current limited development of the corresponding theories.

Returning to the concept of agnosticism, let us consider any possible procedure used to handle observations. Within my agnostic mind, all probability distributions coexist simultaneously, and, in principle, I can analyze one by one the consequences of assuming each of them. Suppose that, for all distributions, it can be concluded that the procedure's outcomes are reliable when applied to an i.i.d. list of observations. Then, I can assert that the procedure is guaranteed solely under a mental stance linking past and future without any preconceptions about how uncertainty behaves at a given point in time. This sheds light on the power of observations and provides a solid footing for inductive reasoning based solely on an indispensable inter-temporal constraint.

§26 An example. We start our discussion with an example.

EXAMPLE 5 (drawing a red ball from an urn) *Consider an urn containing* 100 *balls colored in red and white. I hold that drawing any ball is equally probable, so that the probability of drawing a red ball equals the proportion of red balls in the urn. On the other hand, I have no idea of what the actual proportion is. Consequently, I model the probability of drawing a red ball in the next draw as* $p = \frac{m}{100}$*, where* m *is any number between* 0 *and* 100 *(we have already encountered this setup in §20). Next, I make an experiment: a ball is drawn from the urn, its color is observed and the ball is returned to the urn. This operation is in fact repeated* 1000 *times (which I hold to be independent draws), at the end of which I compute* $\hat{p} = \frac{\#(r)}{1000}$*, the ratio between the number of times a red ball has been observed over the total number of trials. I interpret this ratio as an estimate of the probability* p *of drawing a red ball.*[47]

Figure 4.1 depicts the image probability distribution of $\hat{p}$ *when* $m = 70$*. As it appears, this distribution is concentrated around* $p = 0.7$*. Quite interestingly, a similar concentration result occurs regardless of the value of* m*, and the concentration level of* $\hat{p}$ *around* p *can be rigorously characterized for any value of* m *by means of Hoeffding's inequality:*[48]

$$\mathbb{P}_p^{1000}\{|\hat{p} - p| > \alpha\} \leq 2\exp(-1000 \cdot \alpha^2). \tag{4.1}$$

The interpretation of this formula is as follows. $\mathbb{P}_p$ *is the probability distribution that assigns probability* p *to drawing a red ball. For any given number* α*, the probability of drawing an independent sequence of* 1000 *balls*[49] *such that the estimate* $\hat{p}$ *deviates*

[47] One might correctly argue that the setup described here is not entirely agnostic, in fact p takes one of the values $\frac{m}{100}$ rather than any real value from $[0,1]$. The reason for considering an urn with 100 balls is to give a more concrete appeal to the example (and it also facilitates better graphical representations of the ensuing results). However, all mathematical facts stated in this example remain valid when p can be any real value from $[0,1]$, which indeed corresponds to an agnostic setup.

[48] This inequality was proven by Wassily Hoeffding in [38].

[49] For the notation $\mathbb{P}_p^{1000}$, make reference to Footnote 22.

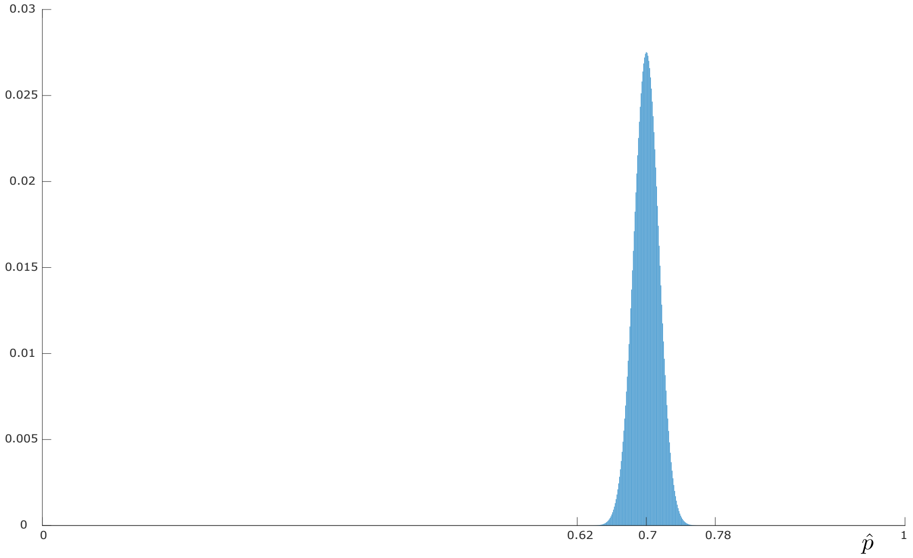

Figure 4.1: Distribution of $\hat{p}$ when $p = 0.7$.

from p more than α *is no more than* $2\exp(-1000 \cdot \alpha^2)$*, twice the exponential function with negative exponent given by* 1000*, the number of trials (so that with increasingly many trials the probability goes to zero exponentially fast), multiplied by* α^2 *(so that taking a small value for* α *increases the value of the bound on the probability). For example, with the choice* $\alpha = 0.08$*, the left-hand side of* (4.1) *yields a value of* $3.23 \cdot 10^{-3}$*, and we can draw the conclusion that* $\hat{p}$ *is an estimate of p within a tolerance of* 0.08 *with (high) probability* $1 - 3.23 \cdot 10^{-3}$ *for all values of p. Figure 4.2 shows graphically this result: for any horizontal line corresponding to a value of p, one sees that the number of red balls that are drawn in* 1000 *trials takes a value so that* $\hat{p} = \frac{\#(r)}{1000}$ *is apart from p no more than* 0.08 *with high probability. From this result we conclude that* $\hat{p}$ *provides valuable knowledge on p without resorting to any prior knowledge (agnostic setup).* $*$

The previous example demonstrates that *probabilistic knowledge can be generated from observations in a fully agnostic setup*. This fact has implications of paramount importance in inductive reasoning that we shall amply explore in this monograph within the framework of learning under consistency. For now, in the next point we will analyze some direct consequences for the example of the urn.

§27 Some consequences. Referring again to Example 5, suppose that we set up a lottery in which an opponent wins one unit of money if the next draw yields a red ball. The composition of the urn is not within our control and, initially, we do not know it; however, we are allowed to conduct experiments on the urn to determine the lottery entry price. To this end, we operate as follows: after drawing a ball 1000 times, the price is set to the value $c = (\hat{p} + 0.08) \cdot (1 - 3.23 \cdot 10^{-3}) + 1 \cdot 3.23 \cdot 10^{-3}$. This entire scheme (i.e., an urn is displayed, we make 1000 draws to determine the lottery price,

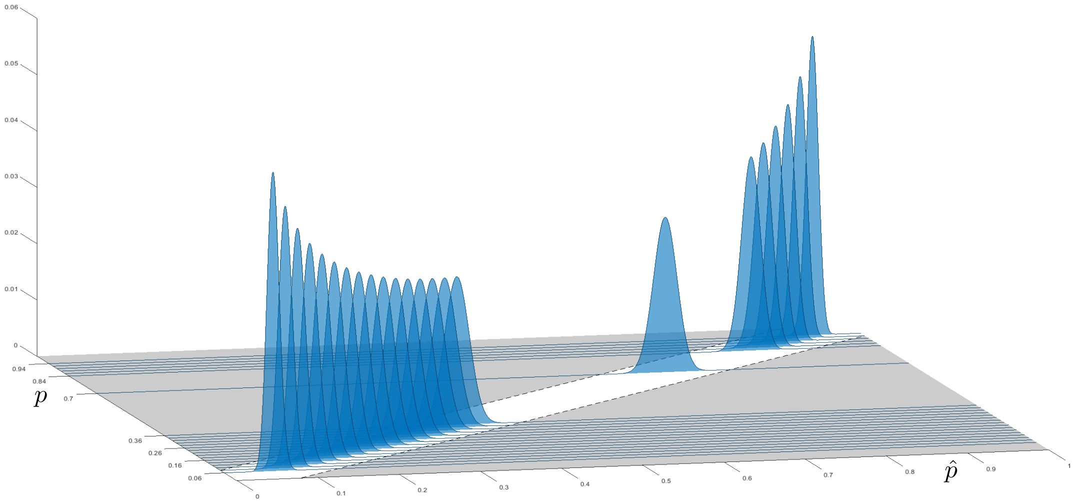

Figure 4.2: Distributions of $\hat{p}$ for different values of p. For each value of p, the white strip is an interval of semi-width 0.08 centered at p.

and then the gambling takes place) is repeated (each time independently of previous instances with a new urn containing an arbitrary proportion of red and white balls) a large number of times. What conclusions can we draw? For any value of p, the expected value of our opponent's winnings is equal to p itself.[50] Also c is a random variable, defined over the set of lists of 1000 draws. Its expected value can be lower bounded as follows: over the lists of draws for which $|\hat{p} - p| \leq 0.08$ (which occurs with a probability of at least $1 - 3.23 \cdot 10^{-3}$), it holds that $\hat{p} + 0.08 \geq p$, while for all other lists we use the trivial bound $1 \geq p$. It then easily follows that $\mathbb{E}[c] \geq p = \mathbb{E}[w]$, that is, the expected value of the price is no less than the expected value of the win. Does this mean that we shall certainly make money in the next bet? Certainly not. However, an application of the law of large numbers[51] reveals that, in the long run, we shall not lose money with probability 1, which is written in formulas as follows:

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} [c_i - w_i] \geq 0 \quad \text{with probability } 1,$$

[50] In elementary probability, the expected value of a random variable f is given by $\mathbb{E}[f] = \sum_{\omega \in \Omega} f(\omega) \cdot \mathbb{P}(\omega)$. This is the mean value of f with probability values acting as weights. In the case at hand, the random variable that describes the winnings – let us denote it with the symbol w – takes the value 1 if a red ball is drawn (which has probability p) and zero otherwise. Hence, $\mathbb{E}[w] = 1 \cdot p + 0 \cdot (1 - p) = p$.

[51] In this monograph, we often refer to classical results from probability theory. In doing so, we exempt us from recalling any time that these results can be found in any textbook on probability. Although referring to classical probabilistic results may seem to be in contradiction with the claim we made in the preface, where we stated that "the probabilistic tools necessary to comprehend the content of this work are introduced, and analyzed in their meaning, within this monograph", we here specify that in the preface we were referring to tools essential for the understanding of the fundamental ideas of this work, rather than specific results in subsidiary parts of our exposition.

where i is an index that runs over lotteries, $c_i - w_i$ gives our net random income (price minus win) on the i-th bet and "with probability 1" means that the result holds over an event that has probability 1 or, which is the same, it can only fail on an event whose probability is zero.[52] Moreover, using results akin to Hoeffding's inequality, one can also establish conclusions that are valid with arbitrarily high probability for an N that is sufficiently large (large, but finite).[53] Of the most importance is that all these results hold without any *a priori* knowledge of the composition of the urns in the various lotteries: information on the composition is only acquired through the experiments.

Since the reader's attention may, at this point, have got trapped into too many technical details, it is important to pause and examine the above result from a distance to appreciate the importance it has in relation to applied fields. For instance, when considering problems in actuarial sciences, by a similar approach one can establish data-driven policies that are probabilistically guaranteed without requiring any *a priori* knowledge on the underlying probability distributions. An insurance company, for example, can set premiums for clients belonging to various groups (categorized by age range, location, educational degree, *et cetera*) by analyzing historical accident records specific to each group (akin to knowledge gained from 1000 extractions in the example of the urn). Likewise, broad are the implications in countless other fields, with the impact of agnostic results growing alongside technological advancements. For example, agnostic results play an increasingly crucial role in machine learning theory and, consequently, in practical applications utilizing machine learning techniques. We shall give more room to various application domains in subsequent chapters.

4.2 Conditional knowledge

§28 Guarantees for a single experimental outcome. Let us take a closer look at the meaning of equation (4.1). It says that observing 1000 balls such that the estimate $\hat{p}$ deviates from p more than α occurs with low probability. This probabilistic statement holds true regardless of the specific value of p. Hence, I can use it when I hold an *a priori* knowledge that restricts the possible values for p, but I can also use it in a fully agnostic setup, as we did in the previous two points.

Now, suppose that, with the data collected in an experiment, I obtain the value $\hat{p} = 0.72$. What can I conclude for this specific value? Let me remark that the importance of this question for a correct understanding of inductive reasoning cannot be overestimated. Suppose I also happen to hold the following sharp *a priori* belief: the

[52] To be precise, one needs to specify that the latter probability distribution refers to an infinite product space whose existence is guaranteed by Ionescu-Tulcea theorem, see, e.g. [61].

[53] Following up what already discussed in §18, we note that all the results (including the limiting result in the above displymath) must only be given a subjectivist interpretation: *if I hold that trials follow an i.i.d. scheme, then, notwithstanding my agnosticism on the composition of the urns, I believe that* .... Surely, this has nothing to do with the frequentist interpretation of probability.

value of p is 0.7. Then, I conclude that my experiments have indeed provided me with an estimate closer to p than $\alpha = 0.08$. On the other hand, if I hold a different sharp *a priori* belief that $p = 0.5$, then I draw the opposite conclusion that I have fallen into the rare event where the estimate is away from p more than $\alpha = 0.08$. From this, we see that there is *no univocal appraisal of the result $\hat{p} = 0.72$ because the evaluation changes depending on extra* a priori *information*. Is this result in contradiction with what we found in §26? Only seemingly. Indeed, a closer inspection reveals that our perspective here has completely changed from §26 in that in §26 we considered all the possible outcomes of the experiments, and our probabilistic statement referred to the overall behavior of the estimation procedure, while here we concentrate on a specific value of the outcome ($\hat{p} = 0.72$). Hence, the two results are not directly comparable. As it often happens, ideas become clear beyond any reasonable doubt by the use of the mathematical language; therefore, in the next point we repeat more precisely the somehow informal reasoning of this point with the help of mathematics.

§ 29 Total vs. conditional probability. Consider again Figure 4.1. It depicts the probability distribution of $\hat{p}$ when $p = 0.7$. This is called the *total probability distribution* when one wants to contrast it with the *conditional probability distribution*, a concept that comes soon after in our discussion. We can notice that the interval $[0.7-0.08, 0.7+0.08]$ contains most of the probabilistic mass. Let us now consider the interval $[0.5, 0.65]$. What is the conditional probability of the event that $\hat{p}$ belongs to $A = [0.7-0.08, 0.7+0.08]$ given that $\hat{p}$ belongs to $C = [0.5, 0.65]$? The answer is provided in §6: it is $\mathbb{P}'_{0.7}(A|C) = \frac{\mathbb{P}'_{0.7}(A \cap C)}{\mathbb{P}'_{0.7}(C)}$ (we use the symbol $\mathbb{P}'_{0.7}$ because this is the image probability distribution of the random variable $\hat{p}$ when $p = 0.7$), which in the present context becomes $\frac{\mathbb{P}'_{0.7}([0.62,0.65])}{\mathbb{P}'_{0.7}([0.5,0.65])} \simeq 0.9999$. According to §16, the interpretation is the following: if we are given the information that $\hat{p}$ is in C, then we should update our subjective belief in the occurrence of an estimate that is no more than 0.08 apart from $p = 0.7$ to the value 0.9999. Graphically, this is the ratio between the area over the interval $[0.62, 0.65]$ and the area over the interval $[0.5, 0.65]$ shown in Figure 4.3. Next, by an inspection of Figure 4.2, we also see that this same reasoning leads to quite disparate values of the conditional probability when p varies. Indeed, the conditional probability of making a correct estimate within a tolerance of 0.08 conditional on the information that $\hat{p} \in [0.5, 0.65]$ spans all the way from 0 to 1.[54] This is the reason why *no meaningful conditional evaluations can be made according to an agnostic approach*. Upon reflection, this fact is not so surprising: it is simply a manifestation of the fact that, in the face of the validity of agnostic evaluations for the total probability, when taking a *more fine-grained point of view* referring to a subset of cases, we can lose our ability to make meaningful assessments. In retrospect, the situation discussed in §28 is nothing but the extreme case when conditioning is taken with respect to one single value of $\hat{p}$ (instead of an interval like $[0.5, 0.65]$), corresponding to full

[54] For those values of p for which $[0.5, 0.65]$ belongs to the interval $[p-0.08, p+0.08]$, the conditional probability is 1, while it drops to zero when $[0.5, 0.65]$ and $[p-0.08, p+0.08]$ do not overlap.

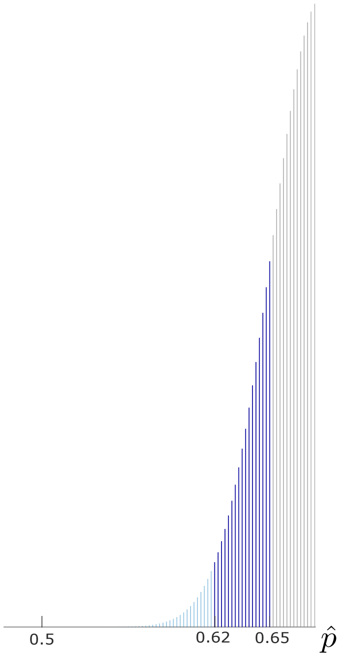

Figure 4.3: Graphical interpretation of the conditional probability $\mathbb{P}'_{0.7}([0.7-0.08, 0.7+0.08]|[0.5,0.65]) = \frac{\mathbb{P}'_{0.7}([0.62,0.65])}{\mathbb{P}'_{0.7}([0.5,0.65])}$: this is the ratio between the dark blue area and the sum of the light and the dark blue areas. The figure is an enlargement of the distribution in Figure 4.1; however, to better emphasize the regions of interest, the drawing in this figure is not in scale.

knowledge of the experimental outcome.

§30 It's all about when one speaks. The previous findings highlight the *importance of when one speaks*. In the example of the urn, speaking after seeing the result of the experiment (as done in §28) is too late a stage to draw any meaningful conclusions in an agnostic setup. However, if one speaks prior to the experiments and considers a more comprehensive standpoint encompassing all possible experimental outcomes (as done in §26), one sees that meaningful evaluations can be formulated. As discussed in §27, in the urn example this matters to set up prices to enter a lottery. More generally, assessing the overall behavior of an experimental procedure is relevant for establishing effective *policies*,[55] with implications in vast domains of control theory, telecommunications, economic sciences, *et cetera*. We shall have opportunities to look into these aspects in later sections of this monograph.

§31 A Bayesian perspective. Referring to a Bayesian perspective may shed further light on the content of this chapter. Suppose that a procedure is proven to return a correct answer in an agnostic setup with some high probability, as it happens in Example 5 where "correct" means that the estimate $\hat{p}$ is within a distance $\alpha = 0.08$ from p and "high probability" is quantified by the value $1-3.23\cdot10^{-3}$. Suppose also

[55] A policy is a rule that indicates the way to operate depending on the observations that have been collected.

that an individual holds an *a priori* probabilistic belief that assigns equal probabilities to each value of p: $\pi(p) = \frac{1}{101}$, for any $p \in \{0, 1, , \ldots, 100\}$.[56] Then, this individual can construct the composition probability distribution $\mathbb{P}$ on the domain of pairs $(p, \hat{p})$ following the approach of §8 and lower bound the probability $\mathbb{P}$ with which the answer is correct. This gives: $\mathbb{P}\{|\hat{p} - p| \leq 0.08\} = \sum_p \mathbb{P}'_p\{|\hat{p} - p| \leq 0.08\} \cdot \pi(p) \geq \sum_p (1 - 3.23 \cdot 10^{-3}) \cdot \pi(p) = 101 \cdot (1 - 3.23 \cdot 10^{-3}) \cdot \frac{1}{101} = 1 - 3.23 \cdot 10^{-3}$. This is a manifestation of the fact that a bound that holds uniformly across all possible cases translates into a numerically equal bound in a Bayesian perspective, and this is true regardless of the individual's prior.

Suppose next that our individual has access to the data from an experiment from which he computes $\hat{p}$ to be 0.72. In force of his prior, he can certainly draw meaningful conditional conclusions. Precisely, $\mathbb{P}(\{|\hat{p} - p| \leq 0.08\} | \{\hat{p} = 0.72\}) = \frac{\sum_{p:|0.72-p| \leq 0.08} \mathbb{P}'_p\{\hat{p}=0.72\} \cdot \pi(p)}{\sum_p \mathbb{P}'_p\{\hat{p}=0.72\} \cdot \pi(p)} = 0.999999996$, which is a fairly high probability. On the other hand, it is interesting to note that this is no way out of the conundrum that no conditional knowledge can be secured in an agnostic setup since a Bayesian individual surely is not agnostic, as he carries his *a priori* belief.

§32 A comment on the philosophical literature. The concepts dealt with in this chapter about when one speaks and about using Bayesian priors to formulate guarantees is well present in the philosophical literature on inductive methods. However, the discussion has often proceeded without an adequate formalization, resulting in slips and misconceptions. Donald C. Williams, [68], draws conditional conclusions without suitable premises to license their validity. David Stove, [63], fallaciously endorses William's thesis. On the other hand, Ian Hacking, [36], lucidly notes that from Williams' premises one cannot infer conclusions that hold for any given sample frequency. We also agree with Patrick Maher, [50], who correctly argues that the conditional step of Williams can only be justified in a Bayesian perspective (even though he does not phrase it this way) by an assumption about *a priori* probabilities "*that is at least as much in need of justification as is induction itself*". This agreeable jumble suggests us that a considerable amount of time and effort might have been saved by using precise mathematical definitions by which one is obliged to stay focused on the correct meaning of concepts.

[56] The probability distribution π over the values of p is called a *Bayesian prior*.

Chapter 5

DECISIONS UNDER OBSERVATIONAL CONSISTENCY

An inductive procedure for constructing models is said to be "observational consistent" if it responds to incorrectly-described incoming observations by invalidating the current model and updating it into a new one. This chapter initially explores modeling procedures grounded in the concept of optimization, and shows that optimization leads naturally to consistency. By this choice, we mean to provide an easy access-point for a concrete understanding of various concepts central to our study. Later, the scope is extended by introducing a more abstract formalization of consistency that moves beyond optimization-based modeling and into decision-making processes.

5.1 Models of a population

§ 33 Models based on observation-constrained optimization. Let $\mathscr{S} = (\omega_1, \omega_2, \ldots, \omega_n)$ be a list of observed members of a population.[57] In this section, we describe a prototypical *procedure* by which a model M^* of the population can be constructed based on $\mathscr{S}$. In §35 we shall see that this procedure is a special case of a general framework for observation-driven decision-making.

Let $\mathscr{M}$ be a collection of subsets of the population, which we interpret as the class of candidate models. In broad terms, our goal is to select a model M^* from $\mathscr{M}$ driven by the following two requisites:

[57] $\mathscr{S}$ will also be referred to as a "sample", which justifies our using the symbol $\mathscr{S}$. The elements ω_i of $\mathscr{S}$ are taken from Ω, the set of observations. Therefore, in the present context Ω corresponds to the population and ω_i are members of the population.

(i) M^* contains the available sample;
(ii) M^* optimizes a quality criterion (typically favoring models able to describe the sample with as little redundancy as possible).

Requisite (ii) expresses a principle of *optimality*, indicating that the model aims to be informative and useful; instead, requisite (i) enforces agreement between the model and those members of the population that have been sampled. It sets *constraints* on the optimization procedure.[58] Before moving to a formal definition, let us swiftly revisit our height and weight Example 1 in §1 to facilitate an intuitive understanding.

EXAMPLE 6 *When a rectangle is employed to describe the height and the weight of the Italian population, as is done in Example 1, requisite (ii) may be taken as the minimization of the area of the rectangle, while (i) prescribes that the points in the sample are within the rectangle.* $*$

Selecting a suitable quality criterion in (ii) is problem-dependent and, in a given application, the choice is often influenced by practical considerations dictated by the intended use of the model. Regardless of the particular choices, a quality criterion as in (ii), along with a class of candidate models $\mathscr{M}$ and the constraints enforced by (i), define a procedure P according to which M^* is selected.

Procedure P

1. input: sample $\mathscr{S}$;
2. optimize with respect to $M \in \mathscr{M}$ the "quality criterion"
   subject to $\omega_i \in M$, for any ω_i in $\mathscr{S}$;
3. output: M^* that solves the optimization program in point 2.

Hence, M^* is the output of procedure P when applied to the sample $\mathscr{S}$, which justifies our using $P(\mathscr{S})$ in place of M^* when we want to be explicit about the sample that has been used.

§34 Examples. We begin with a simple example, well known in the philosophical literature, that of *enumerative induction*.

EXAMPLE 7 (Enumerative induction) *A classical problem in inductive reasoning takes the following form: all objects of type T observed so far have attribute A; what can I conclude about one next object of type T that I shall observe in the future? Will it also have attribute A? For example, all pieces of bread of a certain appearance have thus far been nourishing, can I conclude that a next similar piece of bread will*

[58] In some cases, the model is allowed to fail in representing specific members in the sample that exhibit a "odd" behavior as compared to other members (*outliers*); this results in a smaller model, with improved descriptive capabilities. This situation is part of the framework in §35.

also be nourishing? This is known as the problem of "enumerative induction", see, e.g., Nelson Goodman, [34], and Daniel Steel, [62].

To cast enumerative induction within the framework of our procedure P, let us consider attribute A and its negation $\bar{A}$ (for example: nourish *and* does not nourish*). Given a sample $\mathscr{S}$ consisting of a list of A's and $\bar{A}$'s, P returns the smallest model that contains all cases that have been encountered thus far (consequently, if the sample only contains A, then the model is A itself, whereas encountering also a $\bar{A}$ results in the uninformative model $\{A,\bar{A}\}$ that allows for all possibilities, and observing all $\bar{A}$'s yields the model $\bar{A}$). In other words, the "quality criterion" in procedure P is the total number of cases (A and $\bar{A}$) included in the model, and the selected model is the smallest possible while including all cases that have been observed. As we shall see, the generalization theory presented in the next chapter can be easily applied to this elementary learning scheme.* $\ast$

Our second example is a continuation of Example 1.

EXAMPLE 8 (Chebyshev layer) *Instead of considering a rectangle that contains a sample of Italians as in Example 1, we aim to construct a linear regression model in which the weight is put in relation to the height. Begin by observing that the smallest strip that contains the sample can be constructed by means of the following optimization program:*

$$\begin{aligned} &\min_{\theta_1,\theta_2,\theta_3} \ \theta_3 \\ &\text{subject to: } |weight_i - [\theta_1 + \theta_2 \cdot height_i]| \leq \theta_3, \quad i = 1,\dots,n. \end{aligned} \tag{5.1}$$

Indeed, for given values of θ_1 and θ_2, relation $weight = \theta_1 + \theta_2 \cdot height$ defines a straight line in the $(height, weight)$ domain, called the "central line" of the model. A visualization is provided in Figure 5.1 where $\theta_1 = 6Kg$ and $\theta_2 = 52Kg/m$. Quantity $weight_i - [\theta_1 + \theta_2 \cdot height_i]$ is the vertical displacement of the weight of the i-th individual in the sample from the value taken by the central line corresponding to the height of this same individual. The optimization procedure selects values for θ_1 and θ_2 so that an upper bound θ_3 on the maximum displacement in the sample, always taken as positive thanks to the absolute value $|\cdot|$, is minimized.[59] *Hence, according to program* (5.1)*, all the individuals are "squeezed" into a layer with the smallest possible width θ_3 (which is taken here as the quality criterion to be minimized).*[60]

[59] For the sake of mathematical precision, we note that, if all values $height_i$, $i = 1,\dots,n$, are coincident (which is clearly a peculiar situation), then the solution to (5.1) is not unique (think of tilting the boundaries of the strip as though they were hinged at the two points corresponding to the lowest and to the highest weights). In this case, we assume that a specific solution to (5.1) is singled out by means of an arbitrary *rule of preference* (for example one can decide to select the flat solution with $\theta_2 = 0$).

[60] The construction in this example is a special case of L_∞ regression, see e.g. [37]. It was introduced by Leonhard Euler, [26], some half a century before least squares regression, although a first resolution

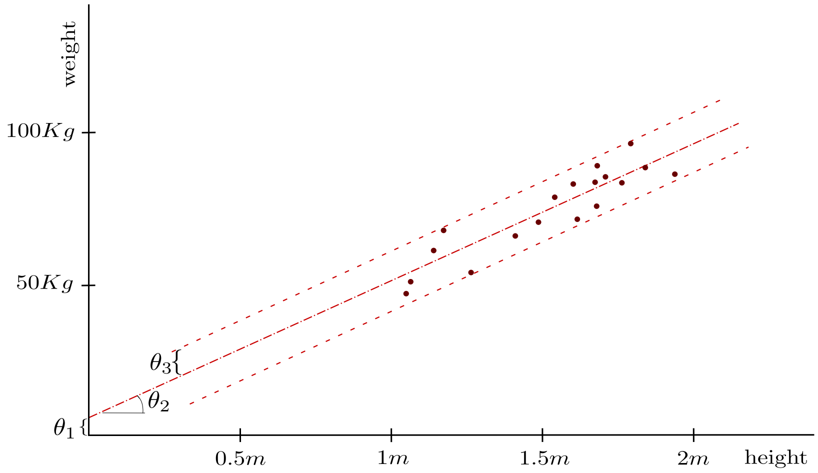

Figure 5.1: Interpretation of the three parameters, θ_i, $i = 1, 2, 3$.

In practice, a model like the one represented in Figure 5.1 can be used for estimation purposes: given the height of a new member of the population, the member's weight is estimated to belong to the line segment at the intersection of the layer with the vertical line corresponding to the member's height. While as futile as it appears in this height-weight example, this scheme gains importance when applied to estimate hidden, difficult-to-measure, features of interest from attributes that are easy to measure. Practical examples include medical applications where the health of a patient is estimated from the outcome of a clinical test. Point §44 presents one such application for the diagnosis of breast tumors. $*$

We next consider the problem of distinguishing objects belonging to two categories.

EXAMPLE 9 (Two classes of objects) *Suppose that the domain $\mathbb{R}^d$ is divided in two half spaces (for example, in $\mathbb{R}^2$ the two half spaces are the regions on either side of a line) containing objects of two different types, denoted as A and B. The location of the hyper-plane that separates the two half-spaces is unknown, but a sample of observations $\mathscr{S} = ((u_1, y_1), (u_2, y_2), \ldots, (u_n, y_n))$, where each u_i is a point in $\mathbb{R}^d$ and $y_i \in \{A, B\}$ is the corresponding object's type, is provided. The task is to derive a model to determine the spatial distribution of objects of type A and B.*

An approach to tackle this problem is as follows. Let $y(u) = w^T u - b$ be a family of linear functions with parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, where $w^T u$ is the scalar product

method for particular cases was provided only in the late 18th century by Pierre S. Laplace, [47]. Since then, L_∞ regression has been considered by various authors, notably by Pafnuty L. Chebyshev and Alfréd Haar, [16, 35], and a layer like the one depicted in Figure 5.1 is at times referred to as a “Chebyshev” layer.

between w and u.[61] *Then, a specific linear function in this family is selected by solving the optimization problem*[62]

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \|w\|^2 \tag{5.2}$$
$$\text{subject to:} \quad \alpha(y_i)\cdot(w^T u_i - b) \geq 1, \ \ i=1,\dots,n,$$

where $\alpha(\cdot)$ is a function that returns -1 if $y_i = A$ and 1 if $y_i = B$. Referring to the top panel in Figure 5.2, function $y^*(u) = (w^*)^T u - b^*$ (with (w^*, b^*) being the solution to problem (5.2)*) is visualized as a slope whose gradient (a directional measure of steepness) is given by w^* while b^* is an offset (changing the value of b shifts the position of the slope). Each observation (u_i, y_i) is represented as a stick of length* 1*, pointing downward (if $y_i = A$) or upward (if $y_i = B$). Constraints $\alpha(y_i)\cdot(w^T u_i - b) \geq 1$ enforce that the slope avoids intersecting the sticks, and minimizing $\|w\|^2$, as done in* (5.2)*, generates the flattest slope among those satisfying these constraints. If we then select the model M^* that assigns A to u when $y^*(u) = (w^*)^T u - b^*$ is below -1 and B when is above* 1 *(refer to the bottom panel of Figure 5.2), this provably generates a model consistent with all the observations while maximizing the distance between the regions labeled as A and B in the model (this distance is called the "margin" in the figure).*[63, 64]

5.2 Decisions with consistency requirements

§35 Procedures with consistency requirements. A model M as in §33 can be seen as a *decision* for the problem of describing a population. However, the concept of decision is broader than this, and it generically refers to any deliberation we make in a given problem. For example, it addresses questions like: how do we allocate a capital across various assets in an investment problem? which therapy is best administered to a patient? or, what maneuvering should be made to avoid a vehicle's collision? Here, we present a comprehensive framework centered around the concept of *consistency*[65] for making decisions based on observations, of which the setup described in the previous §33 is a particular case.

[61] The scalar product between two vectors in $\mathbb{R}^d$, w and u, is given by $w^T u = w_1 \cdot u_1 + \cdots + w_d \cdot u_d$, the sum of the products of the components of equal position in the two vectors.

[62] Symbol $\|w\|^2$ indicates the squared norm of w. It is computed as $\|w\|^2 = w_1^2 + \cdots + w_d^2$, where the w_i's are the component of w.

[63] If all observed objects are of the same type, say $y_i = A$, then the solution to (5.2) remains undetermined: w^* is chosen to be zero, while any $b \geq 1$ satisfies the constraints. However, such an indetermination does not affect the model, which remains the same (the model returns A everywhere) for any feasible choice of b.

[64] Interestingly, the above construction is well-known in the machine learning literature where it underlies a prominent method for classification called Support Vector Machine (SVM), see, e.g., [18, 60, 12]. Classification problems are considered in §36.

[65] This framework has been formulated by Simone Garatti and Marco C. Campi in [29].

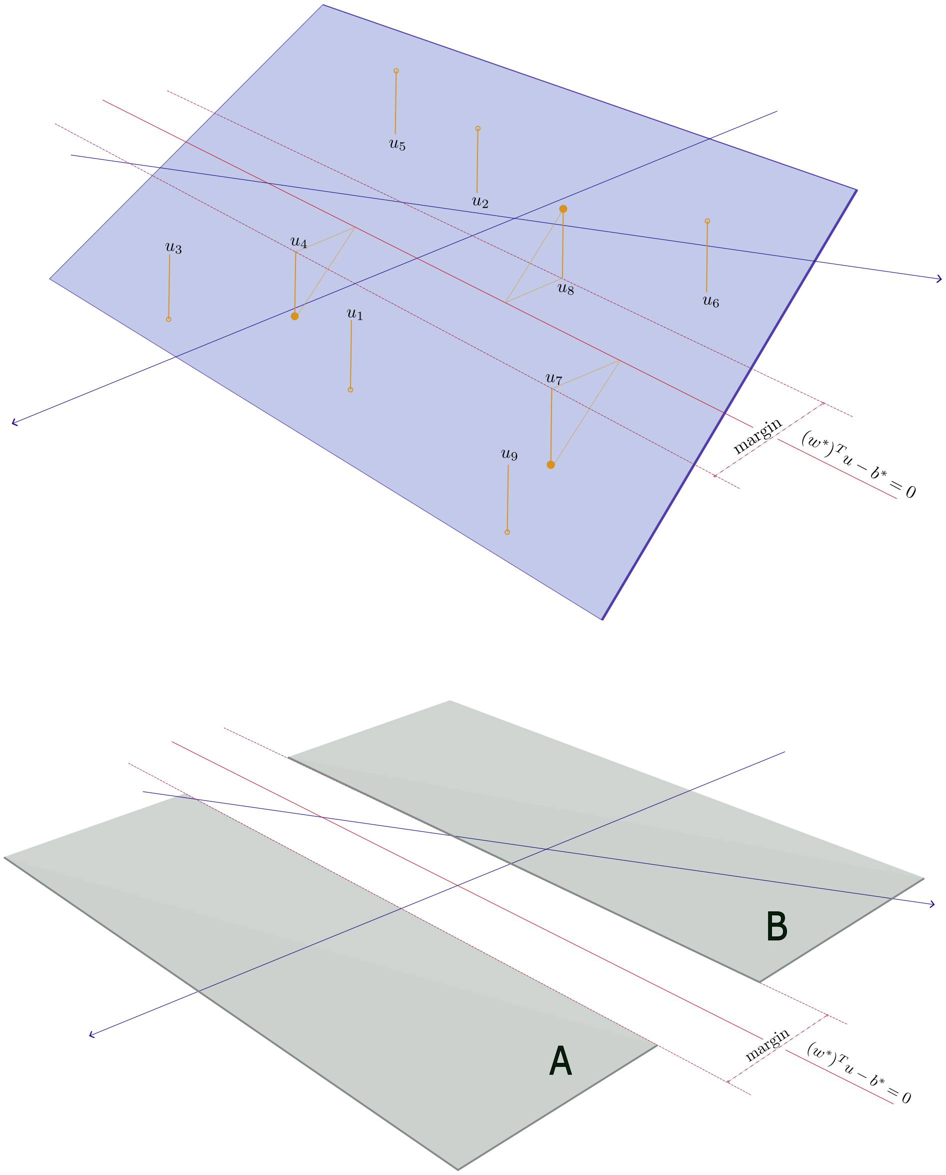

Figure 5.2: *Top*: The linear function selected by problem (5.2). Downward and upward sticks represent the data points with label A and B, respectively. *Bottom*: The model M^*. M^* predicts that objects A and B are located in the corresponding gray regions.

Let $\mathscr{D}$ be a set of decisions. Each ω has associated a subset $\mathscr{D}_\omega \subseteq \mathscr{D}$, the set of decisions that are *appropriate* for ω. For example, in the context of §33 a decision is a model M, and M is appropriate for ω if $\omega \in M$ (hence, $\mathscr{D}_\omega$ is the collection of all models that contain ω, and the requirement (i) can be expressed that the selected model must be appropriate for all observations). Generalizing from §33, we consider procedures P that are maps from a sample of observations $\mathscr{S} = (\omega_1, \omega_2, \dots, \omega_n)$[66] to an element $D^* \in \mathscr{D}$ that satisfy the following assumptions:

(a) **permutation invariance**: for every permutation $(i_1, \dots, i_n)$ of $(1, \dots, n)$, it holds that $P(\omega_1, \omega_2, \dots, \omega_n) = P(\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_n})$ (this means that the order in which observations have been collected is immaterial for the selection of the decision);

(b) **stability in the case of confirmation**: for any $m \geq 1$, if $P(\omega_1, \omega_2, \dots, \omega_n)$ is appropriate for m new observations $\omega_{n+1}, \dots, \omega_{n+m}$ (i.e., $P(\omega_1, \omega_2, \dots, \omega_n) \in \mathscr{D}_{\omega_{n+i}}$ for all $i = 1, \dots, m$), then $P(\omega_1, \omega_2, \dots, \omega_n, \omega_{n+1}, \dots, \omega_{n+m}) = P(\omega_1, \omega_2, \dots, \omega_n)$ (that is, new observations for which the decision is appropriate confirm the decision and leave it unaltered);

(c) **responsiveness to contradiction**: if instead among m new observations $\omega_{n+1}, \dots, \omega_{n+m}$ there is at least one observation ω_{n+i} for which $P(\omega_1, \omega_2, \dots, \omega_n)$ is not appropriate (i.e., $P(\omega_1, \omega_2, \dots, \omega_n) \notin \mathscr{D}_{\omega_{n+i}}$, for some i), then $P(\omega_1, \omega_2, \dots, \omega_n, \omega_{n+1}, \dots, \omega_{n+m}) \neq P(\omega_1, \omega_2, \dots, \omega_n)$ (that is, if the decision is inappropriate even for just one observation, then the decision is changed).

Requirements (b) and (c) are denoted as the *conditions of consistency*. When referencing a procedure that satisfies (a)-(c), we shall often use the term "*consistent procedure*", even though this implies a slight abuse of language since (a)-(c) also include property (a) of permutation invariance.

It is easy to verify that the procedure P in §33 satisfies (a), (b) and (c): optimization is not affected by the order in which observations appear (*permutation invariance*); adding extra observations that are included in the model does not change the optimal solution (*stability in the case of confirmation*); adding even just one new observation that is not in the model forces a change in the solution so as to include the new observation (*responsiveness to contradiction*). On the other hand, in comparison to §33, defining a procedure by the three requirements (a)-(c) introduces a more abstract standpoint that accommodates problems in inductive reasoning beyond model-making via optimization (see §36 for examples).

It is interesting to observe that the conditions of consistency naturally induce *dominance*, broadly meant as the property that some observations are more important, and hence *dominate*, other observations in the process of making a decision. Indeed, suppose that new observations coming downstream in the process of data acquisition do not lead to contradiction (point (c)), so that the decision is maintained (point (b)). Then, these new observations are inconsequential in the formulation of the decision (if they are removed, the decision does not change). Likewise, new observations that change the decision may render previous observations irrelevant to formulate the decision when observations are scanned in a backward fashion. This leads to a condition

[66] The size n of the sample is any non-negative integer and the conditions (a)-(c) below are required to hold for any n.

of dominance: the decision is dictated by only a sub-sample of the observations. Any sub-sample of the observations that returns the same decision as the entire sample is called a *support sub-sample*,[67] and it is not uncommon that support sub-samples are quite restricted as compared to the total amount of observations that have been collected. As we shall see in the next chapter, the concept of support sub-sample plays a prominent role in the reliability theory of inductive reasoning under consistency.

Examples, as presented in the next §36, help understand the generality of the framework introduced in this point.

§36 More examples. The first example is a generalization of Example 2 in §2, it describes a methodology that is often employed in real trading.

EXAMPLE 10 (CVaR) *Conditional Value at Risk (CVaR) is a coherent risk measure in the sense of Philippe Artzner et al., [1], which has been introduced and popularized by R. Tyrrell Rockafellar and Stanislav Uryasev in [58] and [59]. Referring to the investment Example 2, for any θ, let $L_{(i)}(\theta)$, $i = 1, \cdots, n$, be the values attained by $L(\theta, \omega_1), \cdots, L(\theta, \omega_n)$ arranged in descending order: $L_{(1)}(\theta) \geq L_{(2)}(\theta) \geq \cdots \geq L_{(n)}(\theta)$. In statistical terminology, [19], $L_{(n-i+1)}(\theta)$ is called the i-th order statistic of the random sample $L(\theta, \omega_1), \cdots, L(\theta, \omega_n)$.*

Given an integer k in the range $\{1, \dots, n\}$, consider the optimization problem

$$\min_{\theta} \frac{1}{k} \sum_{i=1}^{k} L_{(i)}(\theta). \tag{5.3}$$

When $k = 1$, this comes down to the worst-case approach of Example 2. See instead Figure 5.3 for a graphical representation of the function $\frac{1}{k}\sum_{i=1}^{k} L_{(i)}(\theta)$ when $k = 2$ for the same data set as in Example 2. The reason why in applications one often prefers to take $k > 1$ over the worst-case choice $k = 1$ is that, in this latter case, all the emphasis is placed on just one single (ill) observation and this may result in conservative decisions. In contrast, selecting a $k > 1$ (e.g., in a given proportion to n, for example 5% or 10%) corresponds to minimizing the average of the k worst cases. Intuitively, this allows safeguarding against the occurrence of poor investments while avoiding the over-conservatism inherent in the worst-case approach.

Letting θ^*_{CVaR} be the minimizer of (5.3),[68] define

$$L^*_{CVaR} = \frac{1}{k} \sum_{i=1}^{k} L_{(i)}(\theta^*_{CVaR}), \tag{5.4}$$

[67] The support sub-sample is in general not unique, indeed adding one of the other observations to a support sub-sample gives another support sub-sample. Moreover, there may be two, or more, support sub-samples that are not nested.

[68] It is possible that the minimizer is not unique, which happens if the function minimized in (5.3) is flat along a line of iso-cost; in this case, a θ^*_{CVaR} is singled out by a rule of preference in the domain of θ (e.g., in the case $q = 2$ as in Figure 5.3, one can take the θ^1 with largest value among those that minimize (5.3)).

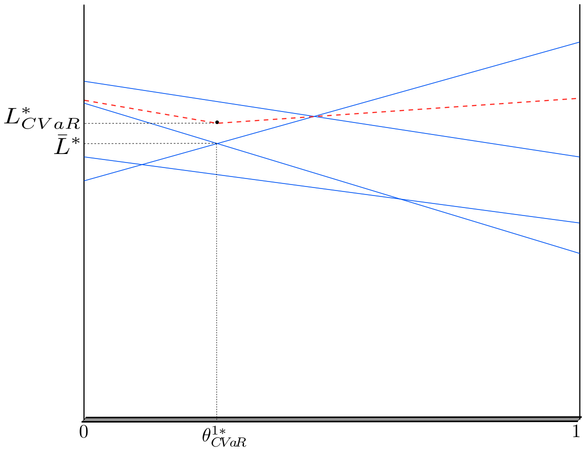

Figure 5.3: Graphical representation of CVaR in the case $q = 2$ assets and $k = 2$. The blue solid lines are $L(\theta, \omega_i)$ and the red dashed line is $\frac{1}{2}[L_{(1)}(\theta^1) + L_{(2)}(\theta^1)]$, the function minimized in (5.3). The value $\bar{L}^*$ is called the "shortfall threshold".

the CVaR value, and

$$\bar{L}^* = L_{(k+q-1)}(\theta^*_{CVaR}), \tag{5.5}$$

which is called the "shortfall threshold" (see again Figure 5.3 for an example). The value of $\bar{L}^*$ represents the tipping point separating shortfalls (tail values bigger than $\bar{L}^*$) from non-shortfalls (values smaller than or equal to $\bar{L}^*$), and it brings in valuable information in addition to L^*_{CVaR} that the user may want to consider when deciding whether or not the solution θ^*_{CVaR} should be accepted and used.[69,70]

If one decides to use the solution θ^*_{CVaR}, then the hope is that a new case ω encountered in the future will not correspond to a shortfall case, i.e., θ^*_{CVaR} will give a performance no worse than the shortfall threshold:

$$L(\theta^*_{CVaR}, \omega) \leq \bar{L}^*. \tag{5.6}$$

This suggests taking as decision the triple $D^* = (\theta^*_{CVaR}, L^*_{CVaR}, \bar{L}^*)$, which also incorporates the shortfall threshold $\bar{L}^*$, and say that it is appropriate for a new ω when (5.6) *holds.*[71]

[69] If the value of $\bar{L}^*$ is deemed unsatisfactory, the user can take various corrective actions, among which increasing the dimension of θ (which means that more assets are included in the portfolio), or even deciding not to operate altogether.

[70] In real applications, the values of L^*_{CVaR} and $\bar{L}^*$ are often complemented with additional indexes such as the average of the empirical values below the shortfall threshold, which provides insight into how well θ^*_{CVaR} performs in non-shortfall cases.

[71] In inductive reasoning, one is interested in evaluating the probability with which a new ω is appropriate. This takes different interpretations depending on the context at hand. For instance, in the context of §33 this means that a new case lies in the model, and therefore it is correctly predicted by the model; in CVaR, it means that ω incurs a loss no more than $\bar{L}^*$, that is, it lies outside the range of shortfalls.

The reader is invited to verify that CVaR satisfies (a)-(c) in §35, in which endeavor one should note that a new loss function $L(\theta, \omega_{n+i})$ that satisfies condition $L(\theta^*_{CVaR}, \omega_{n+i}) \leq \bar{L}^*$ does not expunge the existing decision, while one such that $L(\theta^*_{CVaR}, \omega_{n+i}) > \bar{L}^*$ does prompt a change of the decision: either θ^*_{CVaR} moves to a new location or, if θ^*_{CVaR} remains the same, then L^*_{CVaR} certainly increases.[72,73]

In the next chapter, we shall present generalization results that are applicable to all consistent procedures. Having verified that CVaR is one of them, these findings will enable us to obtain rigorous evaluations of the probability of exceeding the shortfall threshold. Example 22 in §44 will further provide a numerical study on CVaR (borrowed from [54]) that utilizes real data of 10 *assets in the Standard & Poor's S&P500 index.* $\ast$

We next consider binary classification. A *binary classifier* is a predictor that classifies a case described by a vector of attributes into one among two classes, -1 or 1, whose meaning varies depending on the application and can, e.g., be *sick* or *healthy*, *right* or *wrong*, *functioning* or *faulty*. A vector of attributes is called an "instance" and $-1, 1$ are the two possible "labels". For example, in a medical application an instance may contain the outcome of medical tests along with the patient's medical history, and the classifier is employed to determine whether the patient suffers from a particular disease (see §44 for an application to breast tumor diagnosis). Classifiers are often constructed using observations through *machine learning* techniques. An observation consists of a pair (u_i, y_i), where u_i is an observed instance and y_i is the corresponding label; a sample of observations is termed a "training set" (SVM, briefly touched upon in Footnote 64, is a technique used to address this problem). The following example describes a classification technique called Guaranteed Error Machine (GEM).

EXAMPLE 11 (GEM) *The Guaranteed Error Machine is an algorithm for constructing classifiers that has been introduced in [6] and further developed in [14, 5].*[74] *Unlike most techniques used in binary classification, GEM returns a classifier $\hat{y}(\cdot)$ that is permitted to abstain from classifying: $\hat{y}(u) \in \{-1, 1, 0\}$, where issuing the value* 0 *is interpreted as a declaration that the case at hand is too difficult and, hence, the classifier prefers not to provide an answer.*

[72] In the verification of (a)-(c), there is a detail we have glossed over that we feel pressed to at least touch upon here: in CVaR, k in (5.3) has to be regarded as a fixed parameter (if k varies with n, then consistency is easily seen to fail); on the other hand, (a)-(c) must hold for any non-negative value of n (see Footnote 66), which clashes with the fact that n must be greater than or equal to k for (5.3) to make sense (in other words, strictly speaking the procedure is undefined for $n < k$). We advise the reader that this detail is inconsequential, and a comprehensive discussion is given in Section 4.2 of [31].

[73] The reader may have noticed that, in CVaR, D^* is not appropriate for some of the observations $\omega_1, \omega_2, \ldots, \omega_n$, those corresponding to shortfall cases. This is a sign of the fact that the conditions of consistency in §35 do not enforce that the decision be appropriate for all the observations that have been used to formulate the decision.

[74] The algorithm described here is a variation of the one proposed in [5].

Let $\mathscr{S} = ((u_1,y_1),(u_2,y_2),\dots,(u_n,y_n))$ be a sample of observations with $u_i \in \mathbb{R}^d$ and $y_i = y(u_i) \in \{-1,1\}$, where $y(\cdot)$ is an unknown, however complex, two-valued function.[75] *GEM requires the user to choose an integer $k \in \{1,2\dots,n\}$, called the "complexity parameter", which specifies the maximal cardinality of the smallest support sub-sample (refer to §35 for the definition of support sub-sample – we said "the smallest" support sub-sample because there are many support sub-samples and k is an upper bound to the cardinality of at least one of them).*[76] *In loose terms, GEM operates as follows. It is assumed that one has available an additional observation $(\bar{u},\bar{y})$ (in addition to the training set $\mathscr{S}$) that acts as initial "center". GEM constructs the largest possible hyper-sphere*[77] *in $\mathbb{R}^d$ around $\bar{u}$ under the condition that the hyper-sphere does not include any u_i with label y_i different from $\bar{y}$. All points inside this hyper-sphere are assigned the label $\bar{y}$, and all examples (u_i,y_i) for which u_i is inside the hyper-sphere are removed from the training set. The example that lies on the boundary of the hyper-sphere (and that has therefore prevented the hyper-sphere from further enlarging) is then designated as the new center*[78] *and the procedure is repeated by constructing another hyper-sphere around the new center. This time, only the region given by the difference between the newly constructed hyper-sphere and the first hyper-sphere (which has been already classified) is assigned the label of the second center. This iterative procedure continues until either the entire space is classified or the total number of centers reaches the value of the complexity parameter k. In the latter case, the unclassified portion of $\mathbb{R}^d$ is labeled as 0 (see Figure 5.4 for an example of classifier constructed with GEM in $\mathbb{R}^2$).*

This leads to the procedure formally described below.

STEP 0. SET $q = 0$, $S = \mathscr{S}$, $C = \emptyset$ (the empty set) and $u_C = \bar{u}$, $y_C = \bar{y}$;

STEP 1. SET $q = q+1$ and SOLVE problem

$$\begin{aligned} &\max_{r\geq 0} && r \\ &\text{subject to:} && \|u_i - u_C\| \geq r \textit{ for all } (u_i,y_i) \in S \textit{ such that } y_i \neq y_C. \end{aligned} \tag{5.7}$$

Let r^* be the optimal solution (note that r^* can possibly be $+\infty$);

STEP 2. IF $r^* < +\infty$, THEN

[75] In our presentation, it is assumed that the instance is a vector that contains d real variables. More generally, in GEM instances u_i can be elements of a generic Hilbert space, see, e.g., [5].

[76] Choosing a larger value for k reduces the chance of abstention from classifying; when $k > n$, the set of abstention becomes always empty.

[77] This is the same as a sphere in three dimensions (that is, the set of points whose distance from a given point, the center, is no more than a given value), but constructed in $\mathbb{R}^d$ with d any positive number, not necessarily $d = 3$.

[78] In the event of ties, the tie is broken by using an ordering on the points in $\mathbb{R}^d$; for instance, the lexicographic order favoring the u_i that has smallest first coordinate, and, then, the smallest second coordinate if a tie persists, and so forth through all coordinates.

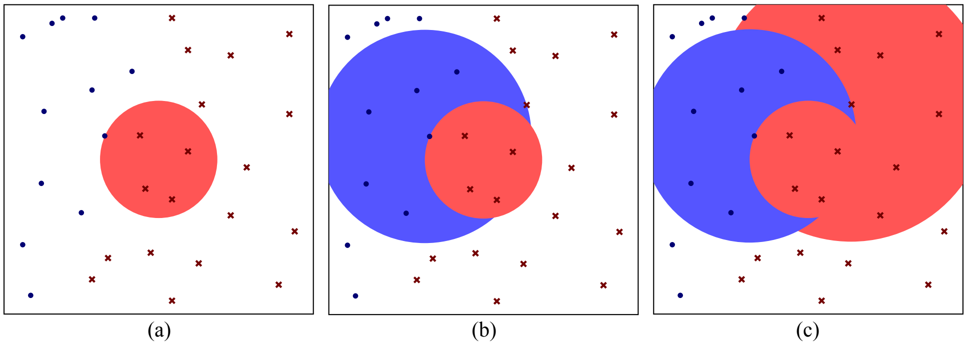

Figure 5.4: A classifier constructed with GEM. The three figures (a), (b), and (c) show the progress of the procedure in costructing the classifier.

2.a *SET* $C = C \cup \{(u_{i^*}, y_{i^*})\}$, *where* (u_{i^*}, y_{i^*}) *is an example in* S *such that: (i)* $\|u_{i^*} - u_C\| = r^*$*; (ii)* $y_{i^*} \neq y_C$*; (iii)* u_{i^*} *is smallest in a given ordering in* $\mathbb{R}^d$ *among all the examples satisfying (i) and (ii);*

2.b *FORM the region* $\mathscr{R}_q = \{u \in \mathbb{R}^d \; : \; \|u - u_C\| < r^* \text{ or } \|u - u_C\| = r^* \text{ and } u \text{ comes before } u_{i^*} \text{ in the ordering in } \mathbb{R}^d\}$[79] *and LET* $\ell_q = y_C$*; UPDATE* S *by removing from it all the examples for which* $u_i \in \mathscr{R}_q$*;*

2.c *SET* $(u_C, y_C) = (u_{i^*}, y_{i^*})$*;*

STEP 3. IF either $|C| = k$ *or* $S = \emptyset$ *THEN STOP and RETURN* $\mathscr{R}_j$, ℓ_j, $j = 1, \dots, q$*; ELSE, GO TO 1.*

The GEM predictor is defined as

$$\hat{y}(u) = \begin{cases} 0, & \text{if } u \notin \mathscr{R}_j \text{ for all } j = 1, \dots, q; \\ \ell_{j^*}, & \text{otherwise, with } j^* = \min\{j \in \{1, \dots, q\} : u \in \mathscr{R}_j\}. \end{cases}$$

As a useful exercise, the reader can verify the validity of (a)-(c) in §35 using the fact that a new observation (u_{n+i}, y_{n+i}) *that is correctly classified* ($\hat{y}(u_{n+i}) = y_{n+i}$) *or that is not classified* ($\hat{y}(u_{n+i}) = 0$) *– in both these cases we say that the classifier is appropriate for the observation – does not change the classifier, while one that is not correctly classified (which corresponds to inappropriateness) does change the classifier. For numerical results with GEM, refer to Example 23.* $*$

Those presented in this point are just two examples to which the data-driven decision making scheme of §35 can be applied. Other examples are found in virtually any domain in which inductive reasoning is applied, including control, power generation and delivery, medical computer-aided diagnosis, regulation of biological systems, to name but a few; the reader may be interested in consulting the position paper [8] that contains an ample presentation of applications in diverse contexts.

[79] The reader may find this definition of $\mathscr{R}_q$ somehow byzantine. The reason why $\mathscr{R}_q$ is defined this way is that it is the easiest to accommodate the conditions of consistency (b) and (c) in §35; see [5] for alternatives.

Chapter 6

COMPLEXITY AND JUDGEMENTS

In Chapter 5, our focus has been on inductive procedures that exhibit a property called "consistency". The primary objective of this chapter is to present generalization results applicable across the wide framework of consistent reasoning. Our voyage will lead us to explore deep-seated mechanisms that link *complexity* (as precisely defined in the chapter) to *judgments* by which inductive reasoning finds a logical foundation. Nonetheless, the process of learning from observations also encounters unchallengeable boundaries, which will also be examined in this chapter.

6.1 Agnostic upper bounds to the risk

§37 Complexity and risk. We introduce two fundamental concepts: *complexity* and *risk*. It turns out that these two concepts are universally linked to each other in inductive reasoning under consistency, as discussed in the next §38.

complexity] Referring to a consistent procedure P as per §35, consider a sample of observations $\mathscr{S}$ and recall that $\mathscr{S}' \subseteq \mathscr{S}$ is a *support sub-sample* if $P(\mathscr{S}') = P(\mathscr{S})$. As previously observed, the support sub-sample need not be unique,[80] and *complexity* indicates the cardinality of the smallest support sub-sample among all. Complexity reflects and quantifies the concept of *dominance*, as previously introduced in §35. Informally, low complexity means that $\mathscr{S}$ can be highly compressed while retaining all necessary information to make the decision. As a case in point, in Example 8 in §34 a support sub-sample is formed by the three individuals that are on the boundary of the Chebyshev layer, while no sub-sample of two individuals returns the same layer, hence the complexity is 3. Note also that, for a given procedure, the complexity depends on the sample at hand (i.e., applying the same procedure to a sample or to another sample may result in different sizes for the smallest support sub-samples in the two cases).

[80] For example, $\mathscr{S}$ itself is a support sub-sample, and various sub-samples of reduced size may exist returning the same decision.

For instance, in the rectangle construction of Example 1 in §1, the complexity will be 4 if four different individuals are the shortest, the highest, the lightest and the heaviest, but it can drop to 3, or even to 2, when one single individual embodies two extreme characteristics (e.g., being the shortest and the lightest at the same time). Since the sample is random, it follows that also the complexity is a random variable.

risk] A model as in §33 is deemed reliable if a new, and yet unseen, observation is contained in the model with high probability. More generally, within the decision context of §35, reliability refers to the probability that a decision is *appropriate* for a new case that will come downstream in the observational flow. This concept is captured quantitatively by the definition of *risk* of a decision D: $R(D) = \mathbb{P}(\omega : D \notin \mathscr{D}_\omega)$. The interpretation of risk is quite diverse depending on the context, and it can be instructive to refer to examples that we have previously encountered to familiarize with it: when regressing the weight of Italians against their height, as is done in Example 8 in §34, the risk is the probability of encountering an individual whose weight is mispredicted by the model on the ground of the individual's height; in the CVaR Example 10 in §36, the risk refers to the probability of incurring a loss higher than the shortfall threshold; and, in a classification problem as in Example 11 in §36, the risk is the probability of misclassification, with manifold implications depending on the problem at hand (e.g., declaring an individual healthy when he is sick, classifying a machine as well-functioning when it is faulty, *et cetera*).[81]

§ 38 Assessing the risk by the complexity. In what follows, we indicate with $D^* = P(\mathscr{S})$ the decision made by procedure P with the sample of observations $\mathscr{S}$ and with c^* the corresponding complexity. The risk $R(D^*)$ depends on the probability distribution $\mathbb{P}$ and, therefore, $R(D^*)$ cannot be univocally determined if multiple determination of $\mathbb{P}$ are deemed possible (see §20). Nevertheless, it has been proven in some contributions (precisely referenced later in this chapter) that $R(D^*)$ can be estimated from the complexity c^* in a fully agnostic setup in which the underlying distribution $\mathbb{P}$ remains unspecified.[82] The importance of this discovery rests in the fact that *c^* can be computed from the observations: in principle, one can test any subsample $\mathscr{S}'$ of $\mathscr{S}$ and verify whether condition $P(\mathscr{S}') = P(\mathscr{S})$ is satisfied; the cardinality of the smallest subsample $\mathscr{S}'$ for which this happens is c^*.*[83] Conceptually, we can trace a parallel between this result and the framework in §26 (see in particular equation (4.1) where the observable $\hat{p}$ is used to estimate p, similarly to using here the observable

[81] The risk is not required to be a very small quantity in all human endeavors. When referring to a fundamental law of physics, it is certainly desirable that the probability with which the law may fail is very small; on the other hand, there are entire fields in financial analytics, medicine, telecommunications where relatively high values of the risk, ranging from a few percent to 10% or even more, are acceptable.

[82] Once more, probability is in my mind and this result says that, even if my mind is fully agnostic, I can draw insightful conclusions on $R(D^*)$ using c^*. Indeed, in an agnostic mind all $\mathbb{P}$'s coexist, and each $\mathbb{P}$ can be inspected in turn; since the result that relates $R(D^*)$ to c^* holds for any $\mathbb{P}$, the conclusion remains valid even without any restriction on $\mathbb{P}$.

[83] See Section 1.2 in [13] for shortcuts to evaluate c^* in the context of non-convex optimization and Section 2 of [11] for convex optimization.

c^* to estimate $R(D^*)$) and the reader is invited to revise the discussion therein for an interpretation of agnostic results, and the subsequent §27 for the consequences thereof.

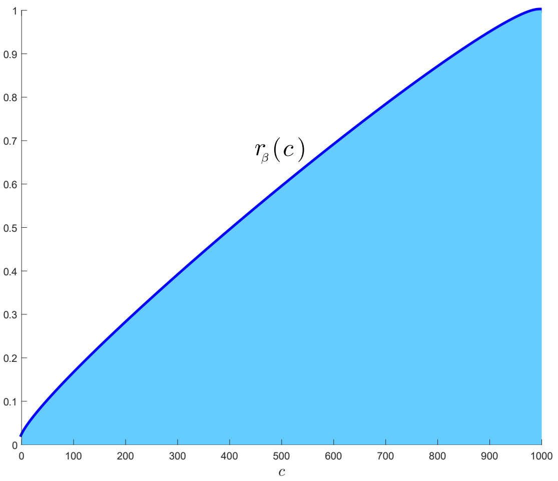

Figure 6.1: Function $r_\beta(c)$ for $\beta = 10^{-6}$ and $n = 1000$. Proposition 12 states that the random pair $(c^*, R(D^*))$ belongs to the blue area below function $r_\beta(c)$ with at least probability $1-\beta = 99.9999\%$, a result that holds for any probability distribution $\mathbb{P}$ (agnostic result).

To state the fundamental result that links complexity to risk, we need to prepare the terrain by introducing a suitable function $r_\beta(c)$ that maps any possible value $c \in \{0,1,\dots,n\}$ of the complexity c^*[84] to a real number in $[0,1]$ that represents an upper limit to the risk (see Figure 6.1 for a representation of $r_\beta(c)$). Function $r_\beta(c)$ also depends on a *confidence parameter* $\beta \in (0,1)$ that the user can freely select (often, β is selected to be a very small value). The interpretation of β is that it sets an upper bound to the probability with which the claim on the risk may fail to be correct, and, conceptually, its meaning is the same as that given to the right-hand side of equation (4.1) in Example 5. To define $r_\beta(c)$, fix a value of $\beta \in (0,1)$ and a value of c in the range $\{0,1,\dots,n-1\}$ ($c = n$ is an exceptional value that needs be treated separately) and consider the following function in the variable $\alpha \in [0,1]$:

$$\Psi(\alpha) = \frac{\beta}{n} \sum_{m=c}^{n-1} \frac{\binom{m}{c}}{\binom{n}{c}} (1-\alpha)^{-(n-m)} ,$$

where n is, as usual, the size of the sample, and symbol $\binom{a}{b}$ indicates the *binomial coefficient*.[85] For any β and c, equation $\Psi(\alpha) = 1$[86] has one and only one solution in

[84] Hence, c^* is the actual value of the complexity, while c is a variable that spans the values that c^* can take.

[85] The binomial coefficient $\binom{a}{b}$ with integers $a \geq b \geq 0$ represents the number of distinct subsets of cardinality b that can be constructed from a set of cardinality a. It turns out that $\binom{a}{b} = \frac{a!}{(a-b)!b!}$. For example, the number of distinct subsets of cardinality $b = 2$ that can be constructed from a set of cardinality $a = 3$ is 3, and this value is given by $\binom{a}{b} = \frac{a!}{(a-b)!b!} = \frac{3 \cdot 2 \cdot 1}{(1)(2 \cdot 1)} = 3$.

[86] While numerical evaluations are not the central focus of attention in this monograph, we never-

the interval $(0,1)$ (see Figure 6.2).[87] Define

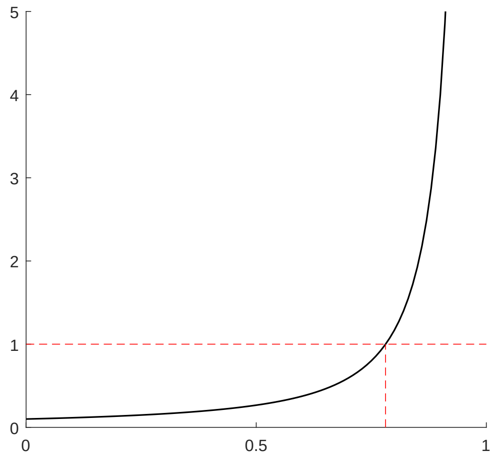

Figure 6.2: Structure of function $\Psi(\alpha)$.

$$r_\beta(c) = \begin{cases} \text{solution to } \Psi(\alpha) = 1, & \text{for } c = 0, 1, \dots, n-1; \\ 1 & \text{for } c = n. \end{cases}$$

Function $r_\beta(c)$ can be shown to be increasing with c. We now have the following proposition.

PROPOSITION 12 (upper bound to the risk) *Consider any consistent procedure P as per §35. Let $\mathscr{S} = (\omega_1, \omega_2, \dots, \omega_n)$ be an i.i.d. (independent and identically distributed) list of observations (see §10) with probability distribution $\mathbb{P}^n$ (see Footnote 22 for this notation) where $\mathbb{P}$, the probability distribution of each single observation, can in my agnostic mind be any distribution. Since D^* depends on the sample $\mathscr{S}$, $R(D^*)$ is a random variable, and so is c^*. For any probability distribution $\mathbb{P}$, the following relation holds:*

$$\mathbb{P}^n\big(R(D^*) \le r_\beta(c^*)\big) \ge 1 - \beta. \tag{6.1}$$

$*$

In equation (6.1), the left-hand side is the probability $\mathbb{P}^n$ of collecting a sample of observations $\mathscr{S}$ for which the risk $R(D^*)$ does not exceed the value given by $r_\beta(c^*)$.

theless notice that a robust MATLAB procedure to solve this equation can be found in Appendix B.1 of [12].

[87] Indeed, $\Psi(\alpha)$ is strictly increasing, continuous, and takes value $\Psi(\alpha) \le \beta < 1$ in $\alpha = 0$ while it grows to $+\infty$ when $\alpha \to 1$.

This probability is lower bounded by quantity $1-\beta$, a value known to the user, and this offers a way to keep control on the probability in the left-hand side of (6.1) that holds true simultaneously for any probability distribution $\mathbb{P}$. While this result may be striking and counter-intuitive, the underlying reason that makes it possible is that changing $\mathbb{P}$ results in a change of both the image probability distribution of $R(D^*)$ and the image probability distribution of $r_\beta(c^*)$. These changes are coordinated in such a way that (6.1) remains valid for any $\mathbb{P}$.[88] This suggests the following practical method to bound the risk: one evaluates c^* (which only depends on the observations) and plug its value into function $r_\beta(c)$; this provides an upper bound $r_\beta(c^*)$ to the risk $R(D^*)$ that holds with high probability $1-\beta$. What is key is that this method has a known (usually high) probability of success with respect to a probability $\mathbb{P}^n$ (which remains unspecified throughout the process) regardless of what this probability is (agnostic result). Rephrasing this fact in looser, but perhaps more "humanized", terms, we can say that: if I use observations to make a decision and indeed I pose no restrictions on how the observations are generated, so reflecting my substantial lack of knowledge (read: $\mathbb{P}$ can be whatever), still I can formulate probabilistic statements on how reliable my decision is (read: how large the risk of inappropriateness is). That is, *a mental stance that just admits the existence of a, otherwise undefined, probabilistic mechanism for the generation of observations licenses the formulation of logically supported statements on the reliability of the decision.* This result applies to any modeling problem as per §33 and, beyond that, to any decision-making problem in the groove of §35.

PROOF OF PROPOSITION 12 [can be skipped without loss of continuity] *Although we have decided not to present proofs in this monograph, and to direct the interested reader to the existing literature for the derivations, in the case at hand we find it necessary to provide some details because this result cannot be found in the literature in the exact form stated here. Nonetheless, it is a relatively brief journey to trace the proof of Proposition 12 back to existing results, a task we undertake in the following.*

Theorem 4 in [5] provides a result that holds for the probability of change of compression for compression schemes that have a preference *property (Property 1 in [5]). To align our setup of decision-making with the theory of [5], we first need to introduce a compression function. This is obtained by compressing a sample (which we see – after removing its ordering – as a multiset*[89]*) into a support sub-sample (again, seen as a multiset – from now on, we do not explicitly remark that samples are identified with multisets) of minimal cardinality. If there exists more than one such support sub-*

[88] Conceptually, this is akin to not knowing $\mathbb{P}_p$ in equation (4.1); however, in contrast to (4.1) the present context covers a truly vast class of inductive methods. In particular, there are no limits to the nature of observations.

[89] A multiset is the same as a set except that it may contain repeats of the same element. For example, $\{1,3,3\}$ is a multiset. Similarly to a set, and differently from a sample, a multiset has no ordering, so that the two multiset $\{1,3,3\}$ and $\{3,3,1\}$ are equal.

samples of minimal cardinality, one is singled out by a rule of preference.[90] *Now, to establish the validity of the preference Property 1 of [5], note that (we use the notations in [5]), if $V = \mathsf{c}(U,z)$ for some z (where c is the "compression function" of [5], and z is an observation ω in our context), then all observations that are in (U,z) and are not in its compression V must be appropriate for $P(V)$ for, otherwise, V could not be a support sub-sample for (U,z) (so contradicting the fact that $V = \mathsf{c}(U,z)$) owing to property (c) of "responsiveness to contradiction". Consequently, either z appears in V as many times as it does in (U,z) (so that $V \nsubseteq U$) or $V \subseteq U$, in which case $P(U) = P(V)$ (owing to property (b) of "stability in the case of confirmation"), so that V is also a support sub-sample for U. On the other hand, $V = \mathsf{c}(U,z)$ also implies that V is minimal and preferred in the rule of preference over any other support sub-sample of (U,z), and so it is* a fortiori *minimal and preferred among support sub-samples of U, implying that $\mathsf{c}(U) = V$. This proves the contrapositive of the condition stated in the preference Property 1 of [5] and, hence, the preference Property 1 is established. To close the rapprochement between our Proposition 12 and Theorem 4 in [5], observe that if an observation is inappropriate for the decision obtained from a sample of observations, then that same observation is inappropriate for the decision obtained from the compression of the sample of observations (because the compression generates the same decision as the sample) and, hence, it gets the compression to change. Therefore, the bound in Theorem 4 in [5] for the change of compression translates into an equal bound for inappropriateness, yielding result* (6.1). $*$

§39 More on probability $\mathbb{P}$. We take a moment to offer a more practical interpretation of the fact that Proposition 12 provides a result that holds irrespective of probability $\mathbb{P}$. Since the result holds for any $\mathbb{P}$, it is applicable in both fully agnostic setups and situations where I hold restrictions on $\mathbb{P}$. In the context of Example 8, $\mathbb{P}$ pertains to the distribution of heights and weights of the population, and I may hold that $\mathbb{P}$ is constrained in specific ways depending on the population. Nevertheless, Proposition 12 stands independently of these constraints, making it applicable to any population. In the CVaR Example 10, $\mathbb{P}$ describes how the vector of rates-of-return distributes. In a "bear market", this vector tends to have lower values than in a "bull market", and the correlation among various components in the vector depends on the composition of the portfolio; for example, assets of similar types (e.g., belonging to the automotive or to the banking sector) are expected to have positive correlation (the so-called "tide effect"). Proposition 12 can be applied to any composition of the portfolio and the nature of the market. Finally, in classification problems, the user may want to include multiple attributes that are deemed useful for the estimation of the label and, correspondingly, $\mathbb{P}$ is a probability distribution that lives in a multidimensional, and possibly highly complex, domain. Proposition 12 applies regardless of this complexity.

[90] A rule of preference just sets an ordering among multisets. A rule of preference has nothing to do with the *preference* property in paper [5], and we apologize for having to use the same word with two distinct meanings.

On the other hand, the fact that Proposition 12 is valid for any $\mathbb{P}$ in no way implies that holding a prior on $\mathbb{P}$ is useless. In modern decision-making problems dealing with complex systems, besides observations, one does want to exploit domain knowledge coming from various sources, often including some that, while not completely trustworthy, can still be of help to obtain a satisfactory solution.[91] It is a fact that all this prior evaluations – alongside with background preferences – can, and should, be used at the time the decision problem is formulated. For example, in the context of portfolio selection with CVaR, leveraging prior evaluations plays an important role when deciding how many, and which, assets should be best included in the portfolio; likewise, in a regression problem as in Example 8, prior evaluations are relevant to decide whether a linear, as opposed, e.g., to a quadratic, center line is best adopted. These choices have an impact on quantitative features of the decision (e.g., the value of the threshold $\bar{L}^*$ in CVaR optimization, or the width of the layer in a regression problem) that the user can directly inspect. What is key is that the rigorous validity of Proposition 12 (which refers to the risk, a quantity that cannot be directly evaluated even after the decision is made) remains intact whether or not the prior that has been used in the problem formulation accurately describes the entire set of distribution $\mathbb{P}$ that are deemed possible. Therefore, by a direct inspection of the decision and the use of the theoretical results, the user comes to "see" the two sides of the medal: (i) features of interest of the decision (through a direct inspection); and (ii) the risk associated with the decision (through application of the agnostic result in Proposition 12 that holds irrespective of the accuracy of any prior the user may have used). The reader may want to refer to the position paper [8], which provides a broad exploration of this topic.

§40 Examples. To gain insight, Proposition 12 is applied to a couple of examples.

EXAMPLE 13 (Charge of an electron) *A sample of electrons is analyzed to ascertain whether their electric charge is positive or negative. If all charges are negative, one makes the model that electrons have negative charge, while observing even one single electron with positive charge, in addition to negative ones, results in an uninformative model that electrons are either negative or positive. In case of all positive charges, the model made is that electrons are positive. This is an instance of "enumerative induction" as described in Example 7 in §34, where T are the electrons and A is the property of having a negative charge.*

Setting, e.g., $\beta = 10^{-6}$ *(such a small probabilistic value is often considered negligible in real applications), with a large set of observations comprising* 400 000 *electrons, the bound in* (6.1) *is* $r_{10^{-6}}(c^*) = 4.94 \cdot 10^{-5}$ *when* $c^* = 1$ *(e.g., all electrons are tested negative).* $*$

EXAMPLE 14 (Coverage of a territory) *Consider a facility, such as a gas station or a laundry, intended to serve the population residing in a specific geographical*

[91] For instance, within an agnostic mind, there may still exist a predisposition to believe that $\mathbb{P}$ is more likely to possess certain characteristics, which one desires to incorporate into the design.

area. To determine a suitable location for this facility, the home locations of 500 *members of the population, drawn at random in an independent fashion, is recorded, and their convex hull*[92] *is used as a descriptor of the population's residential distribution (see Figure 6.3). What is our trust in that one more member of the population, again drawn at random according to the same probability distribution as the other* 500 *members, will indeed happen to live in the convex hull?*

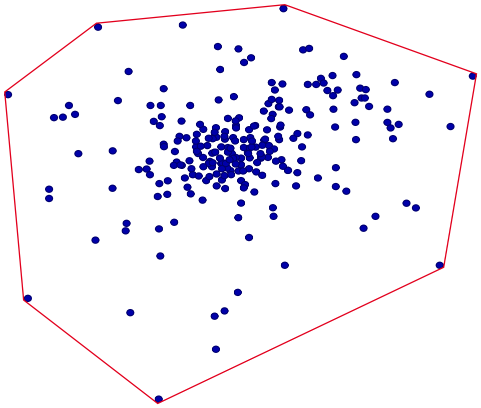

Figure 6.3: Home locations of a sample from the population (blue dots) and the corresponding convex hull (red polygon).

To answer, we observe that the convex hull is a model built according to the method in §33 (which is a particular case of the procedure in §35). Precisely, it is obtained from the procedure:

$$\begin{aligned} \min_{M \in \textit{convex sets in } \mathbb{R}^2} \quad & Area(M) \\ \text{subject to:} \quad & \omega_i \in M, \quad i = 1, \ldots, 500, \end{aligned} \tag{6.2}$$

where ω_i *is a two-dimensional vector that contains the coordinates of the home location of the i-th member in the sample. The risk is the probability that the model is not appropriate for a new member of the population, i.e., the member's home lies outside the convex hull. To apply Proposition 12 to this problem, choose a value for* β*, for example* 10^{-7}*. Then, observing that a support sub-sample for this problem is given by the set of locations that are vertexes of the convex hull (indeed, running problem* (6.2) *with only these locations yields the same convex hull as when all locations are used), one can easily compute the complexity. In the case of Figure 6.3, one finds* $c^* = 7$*, resulting in* $r_{10^{-7}}(c^*) = 7.05\%$*,*[93] *which is interpreted that the proportion of the population residing outside the convex hull is no more than* 7.05%*. Importantly, this result*

[92] A convex set is a set where the line segment connecting any two points in the set is entirely contained in the set. Thus, a square or a disk is convex, but a horseshoe-shaped set is not. The "convex hull" of given points is the smallest convex set that contains all the points.

[93] For computing $r_{10^{-7}}(c^*)$, we used the code referenced in Footnote 86.

holds without resorting to any probabilistic assumption so that we can apply it even when we are fully agnostic on how the territory is populated. $*$

We feel it advisable to remark that it is crucial not to run into a conceptual error when interpreting the result in Example 14. Drawing a parallel with Example 5 in §26, β plays in Example 14 the same role as the right-hand side of equation (4.1) in Example 5. Hence, 10^{-7} is an upper bound to the probability of the following event: a sample of 500 members is drawn, the corresponding convex hull, along with its complexity c^*, are computed, and it happens that the risk associated with the convex hull (which is the probabilistic portion of the population that resides outside the convex hull) exceeds $r_{10^{-7}}(c^*)$. However, this does not have a direct implication on the risk associated to the result for the sample at hand (which is a conditional result – in case of doubts, revise §29. For more on a posteriori assessments in the context of inductive reasoning with consistent rules, see Section 6.3).

§41 Convex optimization. In the context of §33, suppose that the class of candidate models $\mathscr{M}$ is parameterized by a vector $\theta \in \mathbb{R}^d$.[94] The corresponding optimization procedure P is said to be convex if the quality criterion is a convex function of θ[95] and, for any ω_i, constraint $\omega_i \in \mathscr{M}$ restrains θ to be selected from a convex set.[96] It has been shown by Vladimir L. Levin in [49] that, in this convex optimization setup, complexity never exceeds d (the number of components of parameter θ).[97]. Therefore, one knows even before running the optimization procedure that bound $r_\beta(c^*)$ in (6.1) will not exceed $r_\beta(d)$ (recall that function $r_\beta(c)$ is increasing). This *a priori* evaluation can be re-tuned to the value $r_\beta(c^*)$ after c^* has been computed.

EXAMPLE 15 *In the context of Example 1, the rectangle used to describe the Italian population can be parameterized by $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4$, where θ_1 and θ_2 are the coordinates of the center of the rectangle (in the height and weight direction, respectively), θ_3 is the length of the side in the height direction and θ_4 the length of the side in the weight direction. Constructing the rectangle of minimal area that contains the points in the sample amounts to solving the following optimization program:*

$$\min_{\theta_1,\theta_2,\theta_3,\theta_4} \quad \theta_3 \cdot \theta_4$$
$$\text{subject to: } |height_i - \theta_1| \leq \theta_3/2 \quad and \quad |weight_i - \theta_2| \leq \theta_4/2, \quad i = 1,\dots,n,$$

[94] An example of this setup is the construction of the Chebyshev layer in Example 8 where $\theta \in \mathbb{R}^3$.

[95] A function is convex if the line segment between every two points on the graph of the function lies above the graph of the function. For for example, a function shaped like a cup is convex while turning the cup up-side-down results in a non-convex function.

[96] The reader may want to verify that the optimization procedure (5.1) in Example 8 is convex, in which endeavor one has to observe that the inequality $|weight_i - [\theta_1 + \theta_2 \cdot height_i]| \leq \theta_3$ can be re-written by breaking the absolute value in its positive and negative part as follows: $-\theta_3 \leq weight_i - [\theta_1 + \theta_2 \cdot height_i] \leq \theta_3$, resulting in two linear inequalities in θ_i, $i = 1,2,3$. Each inequality defines a half-space in the $(\theta_1, \theta_2, \theta_3)$ domain and the simultaneous verification of the two inequalities holds in the intersection of two half-spaces, which is a convex set.

[97] This fact has been also independently proven, and stated as Theorem 2, in [4].

which can be verified to be convex. Therefore, $r_\beta(c^*)$ is certainly not bigger that $r_\beta(4)$. For example, with $\beta = 10^{-6}$ and $n = 1000$ one obtains $r_{10^{-6}}(4) = 2.7\%$. This a priori *result can be adjusted downwards when one single individual in the sample embodies to extreme characteristics, e.g., being the shortest and the lightest because, as already noticed in §37, this results in $c^* < 4$.* $*$

6.2 Assessments under non-degeneracy

§42 Lower and upper bounds to the risk. No meaningful lower bounds to the risk can be established under the assumptions of Proposition 12. As a case in point, refer to the convex hull in Example 14: if points ω are drawn from a given, finite, set of locations, each with equal probability, then, after drawing 500 points, one may have covered all locations with large enough probability resulting in a risk of zero that one next point will fall outside the convex hull.

Lower bounds can be established under the additional property of *non-degeneracy*.[98] To define non-degeneracy, we need the following additional notion: a support sub-sample is said to be *irreducible* if no observations can be further removed from it without changing the decision (in other words, it is irreducible if does not contain a smaller support sub-sample).

PROPERTY 16 (non-degeneracy) *A learning problem*[99] *is* non-degenerate *if, for any n (recall that n is the sample size), there is with probability* 1[100] *only one irreducible support sub-sample obtained by removing observations from the initial sample in only one way.*[101] $*$

For instance, in the construction of a convex hull as in Example 14, Property 16 rules out the possibility of concentrated masses: if the same point at the vertex of

[98] The terminology “non-degeneracy” has been coined in [9] within the context of convex optimization. In convex optimization, the property of non-degeneracy fails to be true only in situations in which the constraints accumulate in an anomalous way, arguably a “degenerate” condition. However, in other contexts, the term “non-degeneracy” may be somehow inappropriate. We here conform to this terminology to avoid losing an easy link with the relevant literature.

[99] In formal terms, a “learning problem” is defined by a procedure P and a probability $\mathbb{P}$ by which observations are generated.

[100] If we had required the condition to always hold (rather than just with probability 1), this would have led to a property lacking applicability. This is because there is always a chance of observing the same observation more than once, resulting in interchangeable elements within the irreducible support sub-sample. Property 16 implies that such an event occurs with probability zero.

[101] The specification in the final part of Property 16 may seem superfluous; however, it is not: indeed, consider the situation where the initial sample is (a,b,b) and (a,b) is the irreducible support sub-sample. Since (a,b) can be obtained from the initial sample in two different ways (by removing the second or the third element in the sample), this situation violates the property of non-degeneracy.

the convex hull is drawn twice, each of the two draws can be used in forming the irreducible support sub-sample, thereby violating the uniqueness condition in Property 16.[102] We also note that the exclusion of concentrated masses places some $\mathbb{P}$ outside the domain of applicability of the results in this point. Therefore, owing to non-degeneracy, the results in this point are, strictly speaking, not fully agnostic.

To state the proposition that assigns lower and upper bounds to the risk, we need to introduce two functions, $\underline{r}_\beta(c)$ and $\overline{r}_\beta(c)$, where c and β are interpreted similarly to the parameters identified by the same symbols in the function $r_\beta(c)$ of Proposition 12. Fix a value of $\beta \in (0,1)$. For c in the range $\{0,1,\dots,n-1\}$, let

$$\tilde{\Psi}(\alpha) = \frac{\beta}{2n}\sum_{m=c}^{n-1}\frac{\binom{m}{c}}{\binom{n}{c}}(1-\alpha)^{-(n-m)} + \frac{\beta}{6n}\sum_{m=n+1}^{4n}\frac{\binom{m}{c}}{\binom{n}{c}}(1-\alpha)^{m-n},$$

while, for $c=n$, let

$$\tilde{\Psi}(\alpha) = \frac{\beta}{6n}\sum_{m=n+1}^{4n}\binom{m}{n}(1-\alpha)^{m-n}.$$

For any β and $c=0,1,\dots,n-1$, equation $\tilde{\Psi}(\alpha)=1$ admits two and only two solutions in $(-\infty,1)$, say $\underline{\alpha}_c$ and $\overline{\alpha}_c$ with $\underline{\alpha}_c < \frac{c}{n} < \overline{\alpha}_c$[103] (see Figure 6.4). Instead, for $c=n$

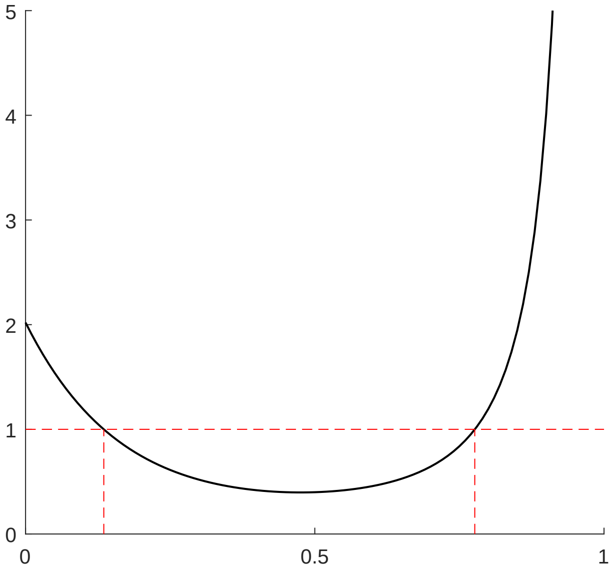

Figure 6.4: (a) Structure of function $\tilde{\Psi}(\alpha)$ for $c=0,1,\dots,n-1$: it tends to $+\infty$ as $\alpha \to 1$ and $\alpha \to -\infty$ and takes a value below 1 in a point in $(-\infty,1)$; equation $\tilde{\Psi}(\alpha)=1$ admits two and only two solutions in $(-\infty,1)$.

[102] One can verify that, in this example, non-degeneracy is indeed equivalent to the assumption of non-concentrated mass, that is, all locations have zero probability of being selected.

[103] For a proof of this fact, see Appendix A in [5].

equation $\tilde{\Psi}(\alpha) = 1$ admits only one solution in $(-\infty, 1)$, which is denoted by $\underline{\alpha}_n$.[104] Define

$$\underline{r}_\beta(c) = \max\{0, \underline{\alpha}_c\}, \quad c = 0, 1, \ldots, n, \tag{6.3}$$

and

$$\overline{r}_\beta(c) = \begin{cases} \overline{\alpha}_c, & c = 0, 1, \ldots, n-1; \\ 1, & c = n. \end{cases} \tag{6.4}$$

PROPOSITION 17 (upper and lower bounds to the risk) *Consider any consistent procedure P as per §35. Let $\mathscr{S} = (\omega_1, \omega_2, \ldots, \omega_n)$ be an i.i.d. list of observations. If the non-degeneracy Property 16 holds, then*

$$\mathbb{P}^n\big(\underline{r}_\beta(c^*) \leq R(D^*) \leq \overline{r}_\beta(c^*)\big) \geq 1 - \beta. \tag{6.5}$$

$*$

The proof of Proposition 6.5 can be found in [29], where it appears as proof of Theorem 2.[105] Figure 6.5 illustrates the result: under the assumptions of Proposition 17,

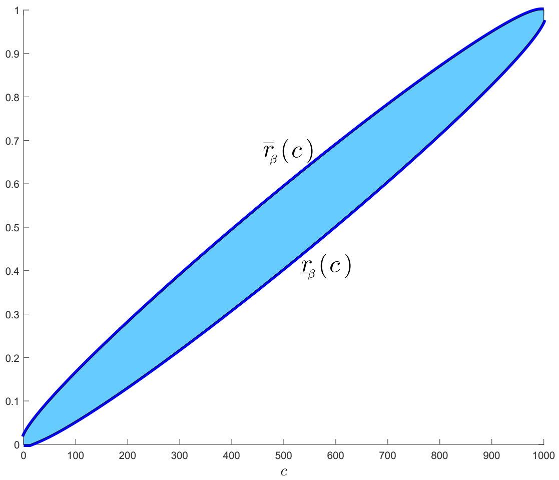

Figure 6.5: Functions $\underline{r}_\beta(c)$ and $\overline{r}_\beta(c)$ for $\beta = 10^{-6}$ and $n = 1000$. In the light of Proposition 17, the random pair $(c^*, R(D^*))$ belongs to the blue elongated area with at least probability $1 - \beta = 99.9999\%$.

the risk is in sandwich between two bounds given by functions $\underline{r}_\beta(c)$ and $\overline{r}_\beta(c)$. Provably[106], these two functions squeeze one on top of the other uniformly in c as the

[104] This is easy to verify because $\tilde{\Psi}(\alpha)$ is strictly decreasing and takes value 0 for $\alpha = 1$ and grows to $+\infty$ as $\alpha \to -\infty$.

[105] Assumption 4 in [29] states the non-degeneracy condition in a form that is different from, but provably equivalent to, the one in Property 16. Proving this fact (note that this requires explicit consideration of the consistency properties) is an interesting, albeit non-trivial, exercise.

[106] See Section 2.1 in [5].

sample size n grows unbounded ($n \to \infty$). In words, this fact is expressed by saying that "the evaluation of the risk is *consistent*". Moreover, the bounds are informative and practically useful for finite values of n. This last point is further discussed in the next example.

EXAMPLE 18 (convex hull in 3 dimensions) *A sample of* 1000 *points is drawn in an independent fashion in* $\mathbb{R}^3$*, and its convex hull is constructed (see Figure 6.6). This problem is non-degenerate provided that points are drawn from a distribution with non-concentrated mass.*

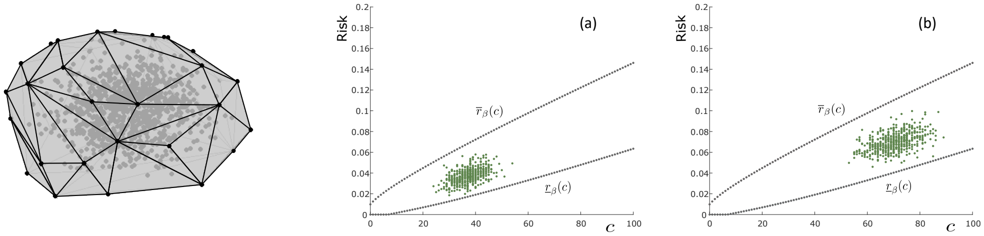

Figure 6.6: Left: convex hull of points in $\mathbb{R}^3$. Right: region delimited by $\underline{r}_\beta(c)$ and $\overline{r}_\beta(c)$ for $n = 1000$ and $\beta = 10^{-3}$. The green dots are generated by a Monte-Carlo testing with the use of the MATLAB procedures for (a) Gaussian distributions and (b) uniform distributions.

Panels (a) and (b) in Figure 6.6 depict the region delimited by $\underline{r}_\beta(c)$ *and* $\overline{r}_\beta(c)$ *for* $n = 1000$ *and* $\beta = 10^{-3}$*. The green dots have coordinates equal to the complexity (horizontal axis) and the risk (vertical axis) in a Monte-Carlo testing in which* 1000 *points in* $\mathbb{R}^3$ *have been generated several times using the MATLAB procedure for Gaussian distributions in panel (a) (this is similar to RandN in Example 4 in §22 but in* 3 *dimensions) and with the MATLAB procedure for uniform distributions in a hyper-cube in panel (b). One sees that the two clouds of green dots in (a) and in (b) are quite different, while, in both cases, they belong to the region. Moreover, the two clouds somehow cover the gap between the lower and the upper bound, which may be interpreted that the bounds are informative. See also §50 for an interpretation of testing using data points.* $*$

§43 An additional bound valid when the complexity is restricted from above. Proposition 17 can be applied without any restriction on the complexity, whose value ranges from 0 to n. When, on the other hand, the complexity admits an upper limit, an additional result on the risk holds. To present this finding, we shall refer to the distribution of the risk alone, instead of jointly considering the risk and the complexity as we did in the previous point. While its practical impact is minor, nonetheless the result of this point holds important epistemological implications, which we shall discuss after the formal statement of Proposition 20.

PROPERTY 19 (limited complexity and problems of complexity d) *A learning problem has* limited complexity *if there is an integer d such that the complexity c^* is no more than d with probability* 1. *The problem has* complexity d *if, for any $n \geq d$, the complexity is equal to d with probability* 1.[107] $*$

Under the non-degeneracy Property 16, all learning problems of complexity d have a universal distribution of the risk (a Beta distribution); in other words, the risk has always the same distribution irrespective of the probability distribution of the observations.[108] Furthermore, problems of complexity d are "worst-case" relative to all problems whose complexity is limited by d. This means that the upper bound on the risk applicable to problems of complexity d also extends to problems with lesser complexity. We first make these results precise, followed by a discussion on their interpretation and significance.

PROPOSITION 20 *Consider any consistent procedure P as per §35. Let $\mathscr{S} = (\omega_1, \omega_2, \ldots, \omega_n)$ be an i.i.d. list of observations. Suppose that the non-degeneracy Property 16 holds and that the problem has limited complexity d. Then, for any $n \geq d$ and $r \in [0,1]$ it holds that*

$$\mathbb{P}^n\big(R(D^*) \leq r\big) \geq 1 - \sum_{i=0}^{d-1} \binom{n}{i} r^i (1-r)^{n-i}; \tag{6.6}$$

moreover, equality holds for all problems of complexity d, that is,

$$\mathbb{P}^n\big(R(D^*) \leq r\big) = 1 - \sum_{i=0}^{d-1} \binom{n}{i} r^i (1-r)^{n-i}. \tag{6.7}$$

$*$

The expression on the right-hand side of (6.6) and (6.7) is a Beta cumulative distribution function with $(d, n-d+1)$ degrees of freedom[109] (see Figure 6.7 for a graphical representation of the density of a Beta distribution). Hence, Proposition 20 can be

[107] The notion of problem with complexity d has been introduced in [9] in the context of convex optimization. As already noticed in §41, in a convex optimization problem in dimension d the complexity never exceeds d. For this reason, in paper [9] a convex optimization problem in dimension d that has complexity d has been named "fully-supported".

[108] For example, the linear regression problem (5.1) of Example 8 in §34 is a non-degenerate problem of complexity 3 whenever the distribution of the observations has a density (a density is a function whose integral over a given region gives the probability of that region, see any textbook on probability theory for a formal definition). Therefore, somehow surprisingly, the distribution of the risk does not depend on the distribution of the observations, provided the latter has a density. Paper [30] contains a complete account of this result.

[109] For any real r, the "cumulative distribution function" of a random variable gives the probability with which the random variable assumes a value less than or equal to r.

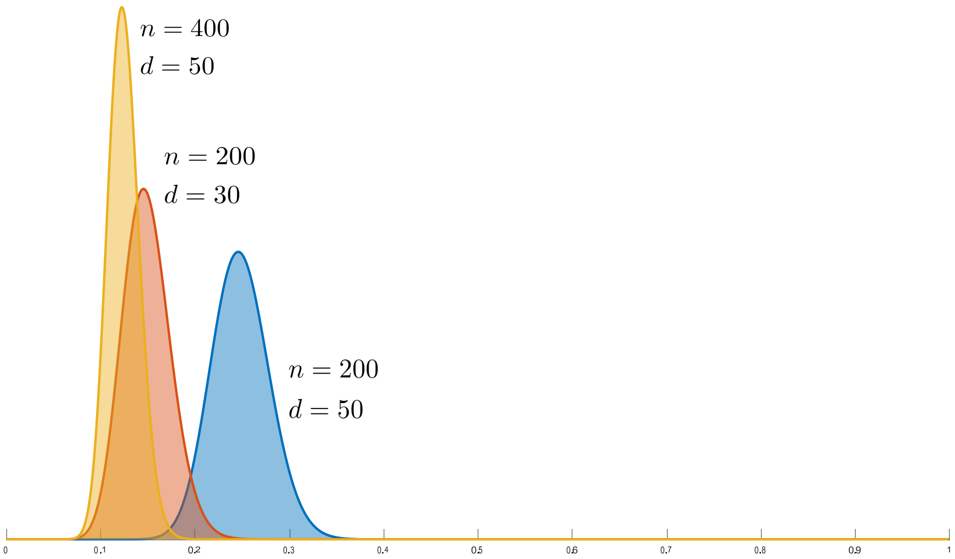

Figure 6.7: Density of a Beta distribution for various values of n and d. One can see that, for $n = 200$, reducing the value of the complexity from $d = 50$ to $d = 30$ gets the density to move to the left (hence, the risk tends to be smaller with lower complexity); a similar effect is obtained by keeping d fixed to the value 50 while increasing the sample size from $n = 200$ to $n = 400$. In the latter case, we also note that the density shrinks, signifying that the probabilistic dispersion of the risk gets less pronounced.

read that the Beta distribution "dominates" the distribution of the risk for problems of limited complexity d (equation (6.6)), while it exactly describes the distribution of the risk for problems of complexity d (equation (6.7)).

PROOF OF PROPOSITION 20 [can be skipped without loss of continuity] *The proof of Proposition 20 is available in the literature for the specific context of convex optimization. Here, we indicate how the existing proof in convex optimization can be generalized to encompass the whole framework of our Proposition 20.*

Result (6.6) *is proven for convex optimization as Corollary 1 in [11]. There, our decision D^* becomes the solution x_N^* and the risk $R(D^*)$ is called the "violation" $V(x_N^*)$. An inspection of [11] reveals that the only properties used to prove Corollary 1 are: (i)* the maximum complexity is d. *While this assumption is not explicitly stated in [11], it comes for granted in the context of convex optimization in dimension d (see §41). In contrast, in the context of our Proposition 20, the bound d to the complexity has been enforced as an assumption; (ii)* the existence and uniqueness of the solution. *The existence and uniqueness of the solution to a convex optimization problem is not always guaranteed; hence, in [11] they are assumed explicitly. In our present context, instead, we consider procedures P that, by definition, return one and only one decision D^*, so that no assumption of existence and uniqueness is needed; (iii)* properties (a)-(c) in §35. *While utilized in the proof of [11], these properties are not assumed explicitly in paper [11] because they hold automatically in the context of convex optimization (as it can be readily verified); (iv)* the non-degeneracy Property 16. *Although* *expressed differently, the Assumption 2 in [11] can be proven to be equivalent to the non-degeneracy Property 16. With the above notices, the proof of Corollary 1 in [11] can be applied* mutatis mutandis *to establish* (6.6) *in Proposition 20.*

As for result (6.7)*, its proof is obtained, again* mutatis mutandis*, from the proof of equation (7) of Theorem 1 in [9], which pertains to convex optimization.* $*$

Comparing Propositions 20 and 12 allows us to highlight some facts that have an interesting interpretation. The right-hand side of (6.7) becomes equal to the right-hand side of (6.1) when $\sum_{i=0}^{d-1}\binom{n}{i}r^i(1-r)^{n-i}=\beta$. Solving this equation for r yields a value that depends on β, d and n, which we write $\tilde{r}_\beta(d)$ (the dependence on n is not highlighted to confirm to other notations). Therefore, (6.7) gives that $\mathbb{P}^n\big(R(D^*)\leq \tilde{r}_\beta(d)\big)=1-\beta$ while, for any $r<\tilde{r}_\beta(d)$, it holds that

$$\mathbb{P}^n\big(R(D^*)\leq r\big)<1-\beta. \tag{6.8}$$

Certainly, $r_\beta(d)$ in Proposition 12 (i.e., $r_\beta(c^*)$ for $c^*=d$) cannot be smaller that $\tilde{r}_\beta(d)$ for, otherwise, applying (6.1) to a problem of complexity d would give

$$\mathbb{P}^n\big(R(D^*)\leq r_\beta(d)<\tilde{r}_\beta(d)\big)\geq 1-\beta,$$

which contradicts (6.8). In fact, computing $r_\beta(d)$ and $\tilde{r}_\beta(d)$ shows that the former is *strictly bigger* than the latter, a disparity that has profound motivations. To comprehend this, consider equation (6.1) in Proposition 12 and, for the sake of the argument (as we shall see, this operation leads to an incorrect result), substitute in it $r_\beta(c)$ with 1 for any $c\neq d$ (so that, when $c^*\neq d$, the proposition gives the void statement $R(D^*)\leq 1$, abstention from providing any meaningful bound to the risk) and $r_\beta(d)$ with $\tilde{r}_\beta(d)$ (so that the risk is claimed to be below $\tilde{r}_\beta(d)$ when $c^*=d$). Then, the left-hand side of (6.1) would become

$$\begin{aligned}&\mathbb{P}^n\big(R(D^*)\leq 1 \text{ and } c^*\neq d\big)+\mathbb{P}^n\big(R(D^*)\leq \tilde{r}_\beta(d) \text{ and } c^*=d\big)\\ &\quad=\ \mathbb{P}^n\big(c^*\neq d\big)+\mathbb{P}^n\big(R(D^*)\leq \tilde{r}_\beta(d) \text{ and } c^*=d\big)\\ &\quad\geq\ 1-\beta.\end{aligned} \tag{6.9}$$

However, a counterexample presented in Appendix 1 of [11] shows that this mathematical claim is false for non-degenerate problems that do not have a bounded complexity d. If we now substitute in (6.9) the bound $R(D^*)\leq 1$ for $c^*\neq d$ with $R(D^*)\leq r_\beta(c^*)$ (as it is in (6.1)), then (6.9) becomes

$$\mathbb{P}^n\big(R(D^*)\leq r_\beta(c^*) \text{ and } c^*\neq d\big)+\mathbb{P}^n\big(R(D^*)\leq \tilde{r}_\beta(d) \text{ and } c^*=d\big)\geq 1-\beta,$$

which – owing to the fact that in this equation we have shrunk the event in the first term on the left-hand side – is even more false than (6.9). Comparing now the last (false) equation with (6.1), we see that, to obtain a mathematically valid result, one has necessarily to lift $\tilde{r}_\beta(d)$ to a higher value, and $r_\beta(d)$ is an appropriate choice, as shown by Proposition 12. The necessity to lift $\tilde{r}_\beta(d)$ carries an interesting interpretation: *when posing a complex question (one that allows for answers whose complexity can exceed d), a posteriori observing that the answer happens to have complexity d does not allow us to draw conclusions on the risk as strong as when the question we ask is itself simple (i.e., the question admits an answer whose complexity is always no more than d).* Or, in more concise form: *answers to simple questions are more guaranteed than simple answers to complex questions.*

Our last result in this point concerns the expected value of $R(D^*)$ (see Footnote 50 for the notion of expected value in the case of elementary probabilities; in the general case, the definition demands some extra mathematical care while retaining the same interpretation as for the elementary case). For problems of complexity d, equation (6.7) says that $R(D^*)$ has a Beta distribution with $(d, n-d+1)$ degrees of freedom. Its expected value is known to be $\frac{d}{n+1}$ (see any textbook in statistics). Instead, the inequality in (6.6) valid for problems with bounded complexity d easily translates into that the expected value of $R(D^*)$ is no more than $\frac{d}{n+1}$. This gives the following result.

PROPOSITION 21 *Under the conditions of Proposition 20,*[110] *if the problem has bounded complexity d, then it holds that*

$$\mathbb{E}\big[R(D^*)\big] \leq \frac{d}{n+1}, \tag{6.10}$$

while equality holds for problems of complexity d, that is,

$$\mathbb{E}\big[R(D^*)\big] = \frac{d}{n+1}. \tag{6.11}$$

$*$

Why are we interested in the expected value? The reason is that the expected value has a useful interpretation: *consider a list* $(\omega_1, \ldots, \omega_n, \omega_{n+1})$ *of* $n+1$ *i.i.d. elements; the expected value is the probability that the first n elements in the list* $\omega_1, \omega_2, \ldots, \omega_n$ *generate a decision that is inappropriate for the last element* ω_{n+1}*:*

$$\mathbb{E}[R(D^*)] = \mathbb{P}^{n+1}\big((\omega_1, \ldots, \omega_n, \omega_{n+1}) : D^*(\omega_1, \ldots, \omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big). \tag{6.12}$$

While the proof of this fact can be found, e.g., in Section 3.2.2 of [10], we demonstrate here its validity for the case of elementary probability.[111]

[110] Since we have presented this result as a consequence of Proposition 20, we require that the assumptions in Proposition 20 hold in the context of the current proposition. However, we feel advisable to notice that the bounded complexity condition is indeed essential for the validity of Proposition 21, while the assumption of non-degeneracy can be omitted without affecting the validity of the results in Proposition 21. To prove this fact, one must follow a demonstrative route that does not make use of Proposition 20 (whose results strictly requires non-degeneracy), and the interested reader can consult [4] for a derivation. An additional reference of interest is the paper [7], which presents this same result in the context of a finite population and exhibits a complete combinatorics-based proof of it.

[111] This derivation can be skipped, and the reader can continue reading from §44 without loss of continuity.

Start by noting that the probability of an event E can be equivalently expressed in one of the following forms: $\mathbb{P}(E) = \sum_{\omega \in E} \mathbb{P}(\omega) = \sum_{\omega} \mathbf{1}(\omega \in E) \cdot \mathbb{P}(\omega)$, where in the last form the summation has been extended to all ω's and $\mathbf{1}(\cdot)$ is "indicator function", which equals 1 when the clause in parenthesis is *true* and it is zero otherwise. We then have

$$\begin{aligned}
\mathbb{E}[R(D^*)] &= \sum_{(\omega_1,\ldots,\omega_n)} R\big(D^*(\omega_1,\ldots,\omega_n)\big)\mathbb{P}^n(\omega_1,\ldots,\omega_n) \\
&= \sum_{(\omega_1,\ldots,\omega_n)} \mathbb{P}\big(\omega_{n+1} : D^*(\omega_1,\ldots,\omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big)\mathbb{P}^n(\omega_1,\ldots,\omega_n) \\
& \quad [\text{where we have used the definition of risk in §37}] \\
&= \sum_{(\omega_1,\ldots,\omega_n)} \Big[\sum_{\omega_{n+1}} \mathbf{1}\big(D^*(\omega_1,\ldots,\omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big) \cdot \mathbb{P}(\omega_{n+1})\Big]\mathbb{P}^n(\omega_1,\ldots,\omega_n) \\
&= \sum_{(\omega_1,\ldots,\omega_n,\omega_{n+1})} \mathbf{1}\big(D^*(\omega_1,\ldots,\omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big)\mathbb{P}^{n+1}(\omega_1,\ldots,\omega_n,\omega_{n+1}) \\
&= \mathbb{P}^{n+1}\big((\omega_1,\ldots,\omega_n,\omega_{n+1}) : D^*(\omega_1,\ldots,\omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big), \qquad (6.13)
\end{aligned}$$

where the last expression coincides with the right-hand side of (6.12).

§44 Two examples. Proposition 21 is illustrated on two examples.

EXAMPLE 22 (average number of shortfalls in CVaR) *We consider the* 5001 *daily closing prices, spanning from November 11, 1995 to October 1, 2015, of* $q = 10$ *companies in the S&P500 index.*[112] *We hold in our mind that the vectors of rates-of-return form an independent sequence*[113] *and that their distribution does not change over the time horizon of interest.*[114] *Conditional Value at Risk – CVaR – is applied in a sliding window fashion (review Example 10 in §36 for the notion of Conditional Value at Risk). Precisely, we consider* $n = 1000$ *consecutive trading days and solve the corresponding CVaR problem, then the window is moved forward by one trading day and CVaR is solved again. At the end, CVaR has been solved* 4000 *times, which we index by* $j = 1,\ldots,4000$. *We hold that the vectors of rates-of-return are sampled from a density in* $\mathbb{R}^q$,[115] *which implies non-degeneracy of the problem.*[116] *In CVaR, k is set to the value* 50, *and one can see that the CVaR problem has complexity*

[112] These companies are those having top market capitalization in the S&P500 at the beginning of 2015, namely, `AAPL`, `XOM`, `MSFT`, `JNJ`, `WMT`, `WFC`, `GE`, `PG`, `JPM`, `CVX`.

[113] Independence of rates-of-return over disjoint periods (e.g., trading days) is a consequence of the Black & Scholes model, which is often adopted in the economics literature. See for example John C. Hull, [39].

[114] Someone might point out that this assumption over a 20 years time frame can be a bit of a stretch, see also §17.

[115] While we hold this assumption as a valid approximation, one might argue that the quantization in prices disrupts it at a fine-grained level.

[116] This fact can be easily proven by considering an irreducible support sub-sample and then observing that one more function $L(\theta,\omega)$ goes through $(\theta^*_{CVaR}, \bar{L}^*)$ with probability zero.

$k+q-1=59$.[117] *All this shows that we are in a position to apply equation* (6.11) *in Proposition 21 and conclude that, for each of the* 4000 *CVaR problems, it holds that* $\mathbb{E}\big[R(D^*)\big] = \frac{k+q-1}{n+1} = \frac{59}{1001} \sim 5.894\%$. *Moreover, in light of* (6.12), *this is the probability that the portfolio selected by CVaR incurs the day after a loss that exceeds the shortfall threshold (*inappropriateness*).*

Now, should the above procedure be applied "with jumps" over non-overlapping intervals of length 1001 *(each interval provides the rates-of-return to construct the CVaR plus one more day for verifying whether the incurred loss exceeds the shortfall threshold) instead of utilizing a sliding window approach, then each interval would only contain variables independent of the variables in other intervals, and the empirical frequency with which the loss exceeds the shortfall threshold would concentrate around value* 5.894% *in the long run (see §18). Although this non-overlap condition is not satisfied in our example due to the one-day shift, still one can argue that this concentration result maintains its validity.*[118] *Figure 6.8 depicts the result for the data set at hand: corresponding to each point j in the abscissa, the blue line gives the average*

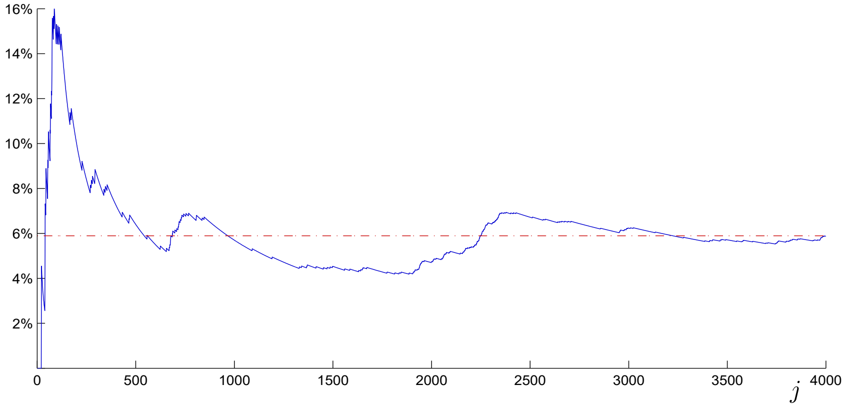

Figure 6.8: Solid blue line (—): for each value of the abscissa, the curve gives the average number of times in which $L(\theta^*_{CVaR,j}, \omega_{j+n})$ exceeds value L^*_j (notice that for, say, $j=1$ CVaR uses $\omega_1,\ldots,\omega_{1000}$, so that the next, test, value is ω_{1001}, which, for generic j, becomes ω_{j+n}; dashed-dotted red line (—·): value 5.894%, obtained from the theory.

[117] Referring to Figure 5.3 where $k=2$ and θ has two components, one sees that, maintaining only the $k+q-1=2+2-1=3$ highest functions corresponding to θ^*_{CVaR}, the decision $(\theta^*_{CVaR}, L^*_{CVaR}, \bar{L}^*)$ does not change; this result is easily extended to generic k and q.

[118] Beyond a concentration result valid in case of a large but finite number of repetitions, even a law of large numbers applies with a one-day shift, and the average of cases in which the loss exceeds the shortfall threshold converges to 5.894%. To show this, it is sufficient to construct many sequences with jumps so as to avoid overlaps (as many sequences as it need be to cover all cases, which is 1001 sequences: the first sequence contains the intervals $[1,1001],[1002,2003],\ldots$; the second sequence the intervals $[2,1002],[1003,2004],\ldots$; $\cdots$; the last sequence the intervals $[1001,2002],[2003,3004],\ldots$) and then observe that the average used to obtain the empirical frequency can be broken up as the average of the averages over the non-overlapping sequences. Since the average over each non-overlapping sequence obeys the law of large numbers, so does the average of the averages.

number of times in which condition $L(\theta^*_{CVaR,j}, \omega_{j+n}) > \bar{L}^*_j$ *is verified on the interval* $[1, j]$. *The dashed-dotted line is at value* 5.894%. $*$

EXAMPLE 23 (breast tumor diagnosis using GEM) *Refer to Example 11 in §36 for an introduction to the Guaranteed Error Machine – GEM.*

Traditionally, the diagnosis of breast tumors has relied on a full biopsy, an invasive surgical procedure. To mitigate the disruption caused by the biopsy, a technique called fine needle aspiration (FNA) has been introduced over the past forty years: a small amount of tissue is aspirated from the tumor, analyzed under a microscope and digitized to finally extract various features of the tumor cells, such as nuclear size, shape and texture. These features serve as inputs to a classifier, which is employed to determine whether the tumor is benign or malignant. Reportedly, however, diagnoses based on FNA are not certain, and FNA is only used as a supplementary tool for the diagnosis, while in doubtful cases a full biopsy remains a necessity, [28, 33].

To construct a classifier for breast tumors, various machine learning algorithms have been proposed, see e.g. [64]. First, a sample of women is tested both with a FNA and a full biopsy so that each woman is described by a set of tumor features and a label, benign or malignant, obtained from the biopsy. This forms the training set, which is used by the algorithm to build a classifier. In a future clinical case, a woman undergoes FNA, and the woman's tumor features are introduced in the classifier to assess the nature of the tumor. Conceptually, this is the same setup as when the weight is estimated from the height with the Chebyshev layer of Example 8 in §34.

We here apply GEM[119] *to a training set of* 683 *cases (*444 *benign and* 239 *malignant) taken from the* UCI Machine Learning Repository*, [2], with* 9 *tumor features, namely,* clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses. *Table 6.1 gives the empirical results. The row with symbol k displays various selections for the complexity parameter of GEM (refer to Example 11 for the definition of "complexity parameter");* $\#(error)$ *is the total number of errors in a* 10*-fold cross validation. Precisely, a window containing the first* 68 *observations (this is the integer part of the total number of observations divided by* 10*) is left out during the training*

k	5	10	15	20	25	30	35
$\#(error)$	5 (0.74%)	12 (1.74%)	20 (2.94%)	25 (3.68%)	26 (3.82%)	29 (4.26%)	31 (4.66%)
$\#(abstension)$	378	347	148	106	49	16	0
$\#(correct)$	297	321	512	549	605	635	649
$k/(n+1)$	0.81%	1.62%	2.44%	3.25%	4.06%	4.87%	5.68%

Table 6.1: Empirical results for the classification of breast tumors.

[119] The numerical results refer to an application of the original GEM algorithm as described in [6].

phase, and the examples in this window are used as tests; the window is then shifted and an adjacent window is used as a test window, and so on ten times until all observations except the last three (that is, $68 \times 10 = 680$ *observations) have played the role of tests. In parentheses, the value* $\frac{\#(error)}{\#(test\ cases)} = \frac{\#(error)}{680}$ *is shown, this is an estimate of the expected value of the risk of the procedure. In the next two rows,* $\#(abstension)$ *and* $\#(correct)$ *are obtained similarly to* $\#(error)$ *summing up the number of times GEM abstains from classifying and the number of correctly classified cases in the* 10*-fold cross validation; the last row gives the bound* $\frac{k}{n+1}$ *for the expected value of the risk as given by equation* (6.10) *in Proposition 21 (recall that, in GEM, k is the largest possible complexity of the classifier; hence, we take* $d = k$ *in Proposition 21), where* $n = 615$ *is the number of training cases (total number of cases,* 683*, minus the observations in the test set,* 68*).*

§45 A digression on exchangeable observations.

We advise the reader that this point is a digression and can be skipped without any loss of continuity. So far, we have considered i.i.d. (independent and identically distributed) lists $(\omega_1, \omega_2, \ldots, \omega_n)$. Interestingly, when $(\omega_1, \omega_2, \ldots, \omega_n)$ is just exchangeable as it was, e.g., in Example 3 in §11, the list can be viewed (with some caveats) as a "mixture of independent and identically distributed lists" (this is known as the "representation theorem"). This fact allows the application of Proposition 21 in the context of exchangeable observations as well, a fact that we briefly discuss in the following.

The representation theorem was originally proven by Bruno de Finetti, [20], in the context of lists that take only two possible values (as is for the Pólya's urn of Example 3) under the condition that the list of observations $(\omega_1, \omega_2, \ldots, \omega_n)$ can be extended to become arbitrarily long without losing the property of exchangeability (this is called "infinite exchangeability").[120] In this context, the interpretation of de Finetti's result is that exchangeable two-valued lists of observations can be viewed as generated by a two-stage process: first, a value p is sampled from $[0,1]$ according to some probability distribution μ and, then, observations are generated i.i.d. with a probability p of drawing one value and $(1-p)$ of drawing the other. Later, this result has been generalized

[120] While this specification may sound bizarre, it is a fact that there are exchangeable lists that do not admit extensions. To illustrate this fact, start by observing that the process employed in the Pólya's urn can continue for an arbitrary number of draws, hence arbitrary extension is possible in this case. On the other hand, the same does not hold for a similar example in which one starts from an urn containing, say, 100 balls, 50 of which are red and 50 white, and the balls are drawn without replacement. In this latter example, it is easy to show that observations are exchangeable for any $n \leq 100$. However, there is no way to make an exchangeable extension to lists of $n > 100$ observations. To see that this is the case, given any list of length 100 that has 50 reds and 50 whites, extend this list by adding a color, red or white, in such a way that the total list of 101 observations has non-zero probability (note that at least one of the two extensions, with a red or a white in position 101, must have a non-zero probability). Now, if we swap the last observation, say that it is a red, with a position where there is a white, then we unbalance the number of reds in the first 100 draws (that is, in the first 100 draws there are now 51 reds), which corresponds to a list of probability zero since in the first 100 draws there must be exactly 50 reds and 50 whites. This rules out exchangeability.

to infinitely exchangeable lists that take value in quite generic domains Ω.[121] When the representation theorem holds, one can "disaggregate" the exchangeable list into its i.i.d. components and apply Proposition 21 to each component. By a process of re-aggregation, it can then be shown that Proposition 21 maintains its validity for the exchangeable list.

As it may have appeared, the presentation in this point was somehow sketchy and served only the purpose of providing a general idea. Then again, we cannot offer a better reading because, at the time of writing this monograph, no technical papers have been yet published on this subject matter.

§46 Popper's theory of conjectures and refutations.

Karl Popper contends that scientific knowledge evolves through *conjectures* and *refutations*: conjectures are formulated by any means and then tested against various observations in an attempt to falsify them. The longer a conjecture survives these attempts of *refutation* the more *corroborated* it becomes, and it is therefore expected to survive new falsification tests as they come along down the stream of observations, [53].

For independent observations, Popper's theory can be put on quantitative grounds by means of Proposition 21. To this end, the first step is to frame the problem within the setup of §33: let us consider a class $\mathscr{M}$ of models that contains only two elements, the conjecture C and a model U, called "universe", that contains all possible observations ω (U corresponds to the void statement that "everything is possible"). C is assigned the smallest value in the "quality criterion" of Procedure P in §33, so that C is selected if all observations ω_i, $i = 1, \dots, n$, are contained in C (that is, observations do not falsify C). In the opposite, C is discarded (*refuted*) in favor of U.

In this context, $\mathbb{E}[R(D^*)]$ in (6.12) can be re-written as follows. If $D^* = C$, then relation $D^*(\omega_1, \dots, \omega_n) \notin \mathscr{D}_{\omega_{n+1}}$ becomes $\omega_{n+1} \notin C$, which means that ω_{n+1} falsifies C; otherwise, when $D^* = U$, condition $D^*(\omega_1, \dots, \omega_n) \notin \mathscr{D}_{\omega_{n+1}}$ is never satisfied because U contains all ω's. Therefore, using result (6.12) we obtain that $\mathbb{E}[R(D^*)] = \mathbb{P}^{n+1}\big(\omega_{n+1} \text{ falsifies } C = D^*(\omega_1, \dots, \omega_n)\big)$. To bound this quantity, we resort to Proposition 21. Noticing that the complexity is zero when C is selected (indeed, when no observations are available, one selects C because it meets the smallest possible value of the "quality criterion"), while it is 1 when C is refuted (in fact, it is enough to keep one single observation outside C to falsify it and generate U), one sees that this problem has bounded complexity equal to 1. Therefore, Proposition 21 gives:[122] $\mathbb{P}^{n+1}\big(\omega_{n+1} \text{ falsifies } C = D^*(\omega_1, \dots, \omega_n)\big) \leq \frac{1}{n+1}$, which shows that the probability of falsifying a conjecture after n confirming tests decreases at a rate (*corroboration rate*)

[121] Still, it is worth noting that the representation theorem has been shown to fail on certain domains Ω that are obtained by rather complex constructions, see e.g. [25].

[122] In fact, the extension of Proposition 21 to degenerate problems, as discussed in Footnote 110. Indeed, this simple reasoning shows that this problem is degenerate: if there are two observations that are not in C, then each of them can serve as an irreducible support list.

that is inversely proportional to n.[123]

Interestingly, Proposition 12 permits one to compute the level of corroboration for procedures that allow on-line updating of the model as new observations come along, so that Popper's conjectures and refutations paradigm becomes a special case of a broader framework.[124] To understand this, suppose that we collect observations in succession and update our model by using a procedure P as per point §33 so obtaining a sequence of models $M_0^*, M_1^*, M_2^*, \ldots$ generated after obtaining $n = 0, 1, 2, \ldots$ observations. If the problem happens to have limited complexity d, then to all these models we can apply Proposition 21 similarly to the theory of conjectures and refutations. In fact, letting $D^*(\omega_1, \ldots, \omega_n) = M_n^*$, the right-had side of (6.12) can be written as $\mathbb{P}^{n+1}\big(\omega_{n+1} \text{ falsifies } M_n^*\big)$, and Proposition 21 bounds this quantity for us with $\frac{d}{n+1}$, again a quantity that decays at a rate inversely proportional to n, as in the conjectures and refutations theory, although with a bigger constant d. When instead there are no upper limits to the complexity, one can conceive the following *modus operandi*: the complexity of the model at hand is measured and elevated to the role of a referee to judge whether the model is too risky to use (in other words, a model is accepted and used only when its complexity does not exceed a barrier, say d, chosen by the user). Studying the risk associated to this setup leads to the conclusion that $\mathbb{P}^{n+1}\big(c^* \leq d \text{ and } \omega_{n+1} \text{ falsifies } M_n^*\big)$ goes to zero at a rate $\ln(n)/n$, not much different from the rate $1/n$ that we found in the case of limited complexity. This result provides a rational justification to the practice of adjusting, even continuously, a model to new empirical evidence provided that the complexity is used to ward off the risk of overadapting the model to the observational data set. Establishing this result requires some additional effort, and we advise the reader that the last part of this point can be skipped without loss of continuity.

Start by considering the random variable $\mathbf{1}(c^* \leq d) \cdot R(D^*)$ ($\mathbf{1}(c^* \leq d)$ equals 1 when $c^* \leq d$ and zero otherwise). It turns out that $\mathbb{E}[\mathbf{1}(c^* \leq d) \cdot R(D^*)] = \mathbb{P}^{n+1}\big((\omega_1, \ldots, \omega_n, \omega_{n+1}) : c^*(\omega_1, \ldots, \omega_n) \leq d \text{ and } D^*(\omega_1, \ldots, \omega_n) \notin \mathscr{D}_{\omega_{n+1}}\big)$,[125] which we re-write for short as $\mathbb{P}^{n+1}\big(c^* \leq d \text{ and } \omega_{n+1} \text{ falsifies } M_n^*\big)$, the probability of measuring a complexity less than or equal to d and the next observation ω_{n+1} is outside M_n^* and thereby falsifies it. To bound $\mathbb{E}[\mathbf{1}(c^* \leq d) \cdot R(D^*)]$, we resort to Proposition 12.[126] This proposition asserts that the lists of observations

[123] It is easy to see that this corroboration rate is tight (not improvable), while the constant 1 can instead be made smaller. Minimizing this constant, however, is not a matter of interest for the purposes of our discussion.

[124] Popper explicitly condemns the practice of adapting theories to observations. Speaking of the Marxist theory of history, in [53] he writes that its followers "re-interpreted both the theory and the evidence in order to make them agree. [...] They thus gave a 'conventionalist twist' to the theory; and by this stratagem they destroyed its much advertised claim of scientific status". We hold, as firmly as mathematics suggests, that adapting theories to observations is possible and scientifically correct provided this process has an impartial judge (the complexity).

[125] In the case of elementary probability, this relation can be proven by a computation similar to (6.13).

[126] While the following argument applies with no restrictions, we invite the reader to concentrate on

$(\omega_1,\dots,\omega_n)$ for which $R(D^*) \leq r_\beta(c^*)$ have probability at least $1-\beta$; considering that the indicator function annihilates the product $\mathbf{1}(c^* \leq d) \cdot R(D^*)$ whenever $c^* > d$ and that $r_\beta(c)$ is an increasing function of c, we then conclude that $\mathbf{1}(c^* \leq d) \cdot R(D^*) \leq r_\beta(d)$ holds at least with probability $1-\beta$. For all other lists of observations (whose probability is no more than β), we use the trivial bound $R(D^*) \leq 1$. Hence, we obtain: $\mathbb{E}[\mathbf{1}(c^* \leq d) \cdot R(D^*)] \leq r_\beta(d) \cdot (1-\beta) + \beta$, a relation that holds for any $\beta \in (0,1)$. Next, we need to bound $r_\beta(d)$. Proposition 8 in [5] does this for us: $r_\beta(d) \leq \frac{d}{n} + 2\frac{\sqrt{d+1}}{n}\left(\sqrt{\ln(d+1)}+4\right) + 2\frac{\sqrt{d+1}\sqrt{\ln\frac{1}{\beta}}}{n} + \frac{\ln\frac{1}{\beta}}{n}$, which, substituted in the previous inequality, gives: $\mathbb{E}[\mathbf{1}(c^* \leq d) \cdot R(D^*)] \leq \left[\frac{d}{n} + 2\frac{\sqrt{d+1}}{n}\left(\sqrt{\ln(d+1)}+4\right) + 2\frac{\sqrt{d+1}\sqrt{\ln\frac{1}{\beta}}}{n} + \frac{\ln\frac{1}{\beta}}{n}\right] \cdot (1-\beta) + \beta$. Selecting $\beta = \frac{1}{n}$ and recalling that $\mathbb{E}[\mathbf{1}(c^* \leq d) \cdot R(D^*)] = \mathbb{P}^{n+1}\big(c^* \leq d \text{ and } \omega_{n+1} \text{ falsifies } M_n^*\big)$, we finally come to the following conclusion: $\mathbb{P}^{n+1}\big(c^* \leq d \text{ and } \omega_{n+1} \text{ falsifies } M_n^*\big) \leq \left[\frac{d}{n} + 2\frac{\sqrt{d+1}}{n}\left(\sqrt{\ln(d+1)}+4\right) + 2\frac{\sqrt{d+1}\sqrt{\ln(n)}}{n} + \frac{\ln(n)}{n}\right]\frac{n-1}{n} + \frac{1}{n}$, where the last expression indeed goes to zero at the rate $\ln(n)/n$.

6.3 A posteriori assessments

§47 The impossibility of conditional assessments. Our discussion here follows that in Section 4.2, points §28 - §30, hence we shall be concise.

Suppose that the charge of 400 000 electrons has been tested negative, while no electrons have been tested positive (refer to Example 13 in §40). What can we conclude about the probability p that one next electron will be positive conditional on these observations?[127] Suppose first that I hold, prior to seen the 400 000 observations, that p is a low number, say $p = 10^{-6}$. Then, I conclude that the observations are in agreement with my holding that p has such a low value. On the other hand, if I deem to know that the risk is high, say $p = 0.1$, then I conclude that my seeing 400000 negative electrons corresponds to a rare event. The very point is that the 400 000 observations alone are not capable of logically excluding one of these two possibilities, both are compatible with the observations and, hence, no agnostic conclusion can be drawn because in an agnostic mind abide simultaneously the two stances that p is as low as 10^{-6} and that $p = 0.1$.

The same conclusion applies similarly to the entire apparatus of inductive reasoning developed in previous sections. Referring, e.g., to Figure 6.1, we know that the probabilistic mass above the blue area is no more than 10^{-6}. Moreover, picking a

the elementary case for a more direct interpretation.

[127] This is the risk associated with making the model that electrons are negative.

value of the complexity c, say $c = \bar{c}$, corresponds to focusing on a vertical line corresponding to that value and, since the whole white region has probability no more than 10^{-6}, then the sole portion of the white region corresponding to $\bar{c}$ must also have probability no more than 10^{-6}. This, however, does not exclude the possibility that also the blue region corresponding to $\bar{c}$ has low probability, so much so that the white portion for $c = \bar{c}$ has probabilistic mass comparable to that of the blue mass, leading to a high conditional probability. Interestingly, if this is the case, we must also conclude that seeing $c = \bar{c}$ corresponds to a rare event.

The interested reader is invited to consult the article [32] for a broad discussion on the impossibility of conditional statements in the context of consistent procedures, as well as the presentation of a Bayesian perspective similar to that we have encountered in §31.

6.4 Drawing the conclusions: justifications and boundaries in the process of learning

§48 Agnostic results. The process of learning from observations, as discussed in previous sections, comprises two foundational components: a *procedure P* that generates decisions based on a list of observations, and the concept of *appropriateness*, which we use to express that one next observation is "in accord" with the decision (e.g., the observation is contained in the model, or the investment does not incur a loss that exceeds the shortfall threshold). Both these components are within our control: we construct the procedure and also define a suitable concept of appropriateness. They formalize the problem of interest as it exists in our mind. In this context, for independent and identically distributed observations, Proposition 12 delivers a distribution-free result, which is a full-fledged demonstration that agnostic evaluations of the risk of inappropriateness are possible. To pinpoint this achievement – perhaps on pain of embracing the somehow vacuous attitude of using catch phrases – we write: *judgments of reliability for inductive procedures are possible in an agnostic setup. Therefore, knowledge can be created out of lack of knowledge in the light of observations.*[128]

In this monograph, an agnostic result was first encountered in §26. Upon critically reviewing the content of §26, we see that its focus can be sided with Proposition 12: the probability p that the next ball is red can be viewed as the risk associated to the

[128] One might argue that a distribution-free result refers to considering any probability distribution on a given support, but the choice of the support is itself an assumption that restricts the way observations are generated. For example, when tossing a coin one can assume that the probability of a head is any number p in the interval $[0,1]$ and that of a tail is $1-p$, thereby excluding that the coin can land in vertical position, which does not correspond to either head or tail. However, in inductive reasoning one is not compelled to consider all available observations, one focuses on observations he is interested in. Hence, outcomes in the vertical position can simply be disregarded and restricting attention to heads and tails is therefore an operative decision, not a restrictive assumption on how observations are generated.

model that claims that the next ball is white. The estimate of this risk is $\hat{p}$. Interestingly, while the result in §26 was simply based on the process of directly estimating p by repeated experiments, Proposition 12 rests on bringing to light a deep-seated connection between two concepts, those of *complexity* and *risk*.

§49 Assessment after seeing the observations. After seeing, say, 400 000 negative electrons without ever observing one single electron with a positive charge, one would be tempted to conclude that, with high probability, the next electron will also be negative. While we have already seen in §47 that conditional assessments (after having seen the observations) are not possible in an agnostic setup, nonetheless we can explore whether conditional conclusions can be justified in some alternative theoretical frameworks. This investigation is certainly worth a bit of our time because it can give us a hand towards a more conscious understanding of important mechanisms that are present in inductive reasoning.

One first attempt of justification is to say that, prior to seeing any of the 400 000 observations, I held that any value of the probability p of coming across of a positive electron is possible and indeed equally likely (this is just an example, and unbalanced judgments are also valid priors). In other words, I held a probabilistic prior that p is uniformly distributed in $[0,1]$ – conceptually, this brings us into a Bayesian framework as in §31. Then, by a computation similar to that used for the conditional probability in the second paragraph of §31, I can compute my conditional belief in that p is, for example, less than 10^{-4} (in other words, while I do not exclude that there exist around positive electrons, still the probability of encountering one is pretty low, no more than 10^{-4}). This gives

$$
\begin{aligned}
&\mathbb{P}\big(p \in [0,10^{-4}] \mid n \text{ negative electrons have been observed}\big)\\
&= \frac{\int_0^{10^{-4}} (1-p)^n \,\mathrm{d}p}{\int_0^1 (1-p)^n \,\mathrm{d}p}\\
&\quad [\text{where } (1-p)^n \text{ is the probability of observing } n \text{ negative electrons}]\\
&= 1-(1-10^{-4})^{n+1}\\
&= \simeq 1-4.24\cdot 10^{-18}\\
&\quad [\text{after substituting } n = 400\,000].
\end{aligned}
$$

This is an extremely strong belief. Interestingly, an individual who holds a completely different prior and assigns a probability as small as 10^{-9} to that p is within the interval $[0,10^{-4}]$, after seeing 400 000 negative electrons, may change completely his mind and come to a substantial agreement with the first individual who holds a uniform prior. For example, assuming that the prior of the second individual has uniform density in the interval $[0,10^{-4}]$ with very small value 10^{-5} (so that the probability that p is in the interval $[0,10^{-4}]$ is $10^{-5}\cdot 10^{-4} = 10^{-9}$) and that the prior also has uniform density in the interval $(10^{-4},1]$ with value $(1-10^{-9})/(1-10^{-4})$, which is close to 1, a computation similar to that made for a uniform prior in $[0,1]$ gives

$\mathbb{P}\big(p \in [0, 10^{-4}] \mid n \text{ negative electrons observed}\big) \simeq 1 - 4.24 \cdot 10^{-13}$.[129]

What lesson can we learn from this? Certainly that holding a probabilistic prior on p justifies the formulation of conditional assessments. On the other hand, while holding a prior is certainly not offensive, one can legitimately investigate more closely where such a prior comes from. One possible answer is that it has nothing to do with experience, it is just an arbitrary act, perhaps dictated by a superior authority one believes in.[130] This is clearly acceptable as it describes one's beliefs. However, we – as people interested in inductive reasoning – have to make clear that here *our judgements have been licensed by an extra element, besides observations, that have played a crucial role in the formulation of our final conclusion*.[131] On the other hand, an alternative answer – indeed often advocated by many – is that the prior encapsulates, and summarizes, previous experience. This standpoint, however, must be carefully analyzed to avoid misconceptions. Previous experience may perhaps refer to a "super-experiment": at some point in time, I have had at my disposal a machinery able to scrutinize all electrons in the universe, and it turned out that the charge of them all was negative. From this knowledge, I can certainly form my prior that $p = 0$ has probability 1. If instead previous experience merely refers to the assessment of the electron charge in multiple experiments that have preceded our 400 000 observations, we should refrain from drawing a prior simply because this prior would be a conditional probability based on partial experience acquired in the past, which leads to a conundrum as we know this is not possible without a prior.

To get around of the above difficulty, one suggestion that has been made is that, after removing all experience (the 400 000 observations and all experience that has preceded these observations) one gives a uniform prior on p according to the "principle of indifference" (see §14).[132] While I might agree that this is reasonable, someone else (perhaps myself in another state of mind) can be against it. Hence, again, we

[129] We interpret this result that being exposed to a common empirical evidence has a strong effect on narrowing the initial distance among people's opinions.

[130] Religions, for example, provide prior beliefs.

[131] It is worth noting that if my prior on p assigns zero probability to the interval $[0, 10^{-4}]$, then the conditional probability of $[0, 10^{-4}]$ after seeing 400 000 negative electrons remains zero, in complete disagreement with the two individuals whose opinion has been discussed before. Hence, what we have called the "extra element" is able to steer our judgments from one extreme to the other.

[132] When p is assigned a uniform prior, the conditional probability of any interval $[\underline{p}, \overline{p}]$ of length Δ is given by $\mathbb{P}\big(\{p \in [\underline{p}, \overline{p}]\} \mid \{\text{empirical evidence}\}\big) = \frac{\mathbb{P}(\{p \in [\underline{p}, \overline{p}]\} \cap \{\text{empirical evidence}\})}{\mathbb{P}\{\text{empirical evidence}\}} = \mathbb{P}\big(\{\text{empirical evidence}\} \mid \{p \in [\underline{p}, \overline{p}]\}\big) \cdot \frac{\mathbb{P}(p \in [\underline{p}, \overline{p}])}{\mathbb{P}\{\text{empirical evidence}\}} = \mathbb{P}\big(\{\text{empirical evidence}\} \mid \{p \in [\underline{p}, \overline{p}]\}\big) \cdot \frac{\Delta}{\mathbb{P}\{\text{empirical evidence}\}}$. Function $\mathbb{P}\big(\{\text{empirical evidence}\} \mid \{p \in [\underline{p}, \overline{p}]\}\big)$ is called the "likelihood" of the the empirical evidence when p falls within the interval $[\underline{p}, \overline{p}]$, and we see from the last formula that the conditional probability of any interval $[\underline{p}, \overline{p}]$ of length Δ given the empirical evidence is proportional to the likelihood of the empirical evidence when p is in the considered interval (this is because Δ and $\mathbb{P}\{\text{empirical evidence}\}$ are constants). This observation relates to common judgments of the type: "I don't believe that p is in the interval $[\underline{p}, \overline{p}]$ because this would make the $\{\text{empirical evidence}\}$ I have had quite unlikely." This sentence refers to the probability of $\{\text{empirical evidence}\}$ given $[\underline{p}, \overline{p}]$

have to register that *an extra element, besides observations, plays a central role in the formulation of a judgment grounded on the principle of indifference*. Said differently, the principle of indifference is nothing but one more way to instate a prior that is not dictated by observations.

Summarizing, we come to the following apodictic conclusion: *if a super-experiment is not available, no conditional beliefs can be formed only based on observations; any conditional belief has necessarily to be supported by additional elements beyond the observational wealth. This is what we call the "inescapable relativism of conditional beliefs".*

§50 Test of hypothesis. Suppose that we hypothesize a probabilistic model, and that this model logically implies that the observations have a given behavior B with known, high, probability $1-\gamma$. We also decide that, upon collecting the observations, we shall accept the model if indeed the observations will exhibit the behavior B. This is called a *test of hypothesis*. Clearly, the probability of rejecting the model when it is correct has probability $1-\gamma$.[133] On the other hand, in line with the discussion in §49, conditional conclusions after having actually seen the observations are not possible under general circumstances and, hence, the above probabilistic guarantee refers to the method, not to the specific conclusion that is obtained in a single application of it. It has to be further noticed that, in practical usage, this procedure is often applied in a "soft manner" in which the implications of the assumed model are not fully specified at the time the method is applied. This is the perspective in which the empirical testings with real data provided in various parts of this monograph should be interpreted.

(which is the likelihood) and not to the probability of $[\underline{p},\overline{p}]$ given {empirical evidence} (which is what we are interested in). Even though this is often not consciously observed, the link between the two passes through the use of a prior (which, in our case, is expressed by setting $\mathbb{P}(\{p \in [\underline{p},\overline{p}]) = \Delta$, a uniform prior).

[133] Instead, it is impossible to quantify the probability of accepting the model when it is incorrect, this would require a description of all the alternatives.

The Epilogue

Inductive reasoning is all about forming beliefs guided by observations. Observations are the input to this process and its output, the beliefs, are important to us as they represent our knowledge and guide us in the deliberations we make to decide how to act on the world we perceive.

Our treatment of inductive reasoning has been minimalist. Observations exist insofar as they are perceived, and our perceived observations are processed by rules that we prescribe. These rules are not meant to be descriptive, if not just descriptive of our own choices. Our study does not require any additional elements beyond observations and the rules used to process them.

Inductive reasoning is a one-directional process: from observations to beliefs. This, however, does not exclude the possibility of using observed facts, about which we have previously made a prediction, to adjust our inductive rules. For instance, we might create various prediction schemes using a first set of observations and then test these schemes against additional observed facts for which we have formulated a prediction. The rule that scores higher in terms of prediction capabilities will then be selected as our rule. Or, perhaps, we shall adopt an average of the best rules. What matters is that, after the additional facts are observed, they become part of our observational wealth, and the entire process formed by the method used to construct the prediction schemes, augmented with the selection criterion based on the scores on subsequent observations, can be interpreted as a big function that maps the total set of observations into our beliefs. Using this function to furnish a prediction on an as-yet-unobserved fact is again a one-directional operation.

While inductive reasoning is one-directional, some individual may also be willing to introduce an “anchor” to attest the quality and effectiveness of a given inductive approach before actually observing how the forecasted facts unfold. This requires setting up a framework in which one posits the existence of the real world, and assigning laws that govern its evolution. But this individual is not us. Positing and describing the real world is too much of a leap of faith for us to accept it. Indeed, ascertaining the existence of the real world in any given form cannot be the outcome of an inductive process, simply because induction refers to what we think, and all we can think of are our own thoughts. The real world is not accessible as such, and we refrain from positing its existence, which would give a mystical bent to our treatment of inductive reasoning. Moreover, to someone willing to introduce the real world, an attitude we certainly do not condemn, we would suggest to first adopt a minimalist stance and see how far he can go with it. This would target a program of clarity, in the attempt to identify what premises are indeed necessary to draw specific conclusions.

Inductive reasoning cannot subsist without prescribing some rule that links past to future, what has been seen to the unseen. In this monograph, we have pervasively used the assumption of independence and identical distribution (i.i.d.). This assumption has not to be universally accepted: it will be put at the service of our inductive process when we hold it applies to the context at hand. Crucially, under all circumstances, the i.i.d. assumption is prescriptive, and we will hold in high regard the consequences that stem from it to the extent that this assumption has been introduced with honesty, truly believing it is apt for the context in which we operate.

Someone can regard the i.i.d. assumption as being demanding, an opinion we concur with. Its adoption is solely motivated by the technical fact that, at the present time, no alternatives exist able to provide results having the same depth and penetration as those presented herein in the i.i.d. setup. While further developments can be expected to ease this assumption, what remains is our surprise, which we expect to be our reader's surprise, for the impressive strength of the results that can be obtained under only assuming an i.i.d. framework beyond any further description of the underlying probabilities. These results show that a subjective rule linking the past to the future, the seen to the unseen, suffices to justify the use of inductive reasoning without any further prejudgments on the perceived world.

Since our results are distribution-free, they hold for all distributions and, thereby, for any subset of distributions. This offers a rational ground on which any person, independently of the subjective distributional prior, can accept the guidance of inductive reasoning to make deliberations. However, this does not resolve the issue that, upon seeing how facts unfold, someone may find them poorly aligned with his prior, leading to the conclusion that the observations have fallen in a rare event. This is a manifestation of the impossibility to establish an inductive theory for *a posteriori* assessments that has universal validity.

Bibliography

[1] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[2] A. Asuncion and D.J. Newman. UCI machine learning repository, URL: http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences. 2007.

[3] P. Berti, E. Miranda, and P. Rigo. Basic ideas underlying conglomerability and disintegrability. *International Journal of Approximate Reasoning*, 88:387–400, 2017.

[4] G.C. Calafiore and M.C. Campi. Uncertain convex programs: Randomized solutions and confidence levels. *Mathematical Programming*, 102(1), 2005.

[5] M. C. Campi and S. Garatti. Compression, generalization and learning. *Journal of Machine Learning Research*, 24:1–74, 2023.

[6] M.C. Campi. Classification with guaranteed probability of error. *Machine Learning*, 80:63–84, 2010.

[7] M.C. Campi. Inductive knowledge under dominance. *Synthese*, 201:1–29, 2023.

[8] M.C. Campi, A Carè, and S Garatti. The scenario approach: a tool at the service of data-driven decision making. *Annual Reviews in Control*, 52:1–17, 2021.

[9] M.C. Campi and S. Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[10] M.C. Campi and S. Garatti. *Introduction to the scenario approach.* MOS-SIAM Series on Optimization, 2018.

[11] M.C. Campi and S. Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018.

[12] M.C. Campi and S. Garatti. A theory of the risk for optimization with relaxation and its application to support vector machines. *Journal of Machine Learning Research*, 22(288):1–38, 2021.

[13] M.C. Campi, S. Garatti, and F.A. Ramponi. A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control*, 63:4067–4078, 2018.

[14] A. Caré, F.A. Ramponi, and M.C. Campi. A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Systems Letters*, 2:393–398, 2018.

[15] R. Carnap. *Logical foundations of probability*. The University of Chicago Press, 1950.

[16] P.L. Chebyshev. *Mémoires présentés a l'Académie Impériale des Sciences de St. Pétersbourg par divers savants*, 7:539–568, 1854.

[17] D.M. Cifarelli and E. Regazzini. De Finetti's contribution to probability and statistics. *Statistical Science*, 11 (4):253–282, 1996.

[18] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[19] H. A. David and H. N. Nagaraja. *Order statistics, 3rd. ed.* Wiley, 2003.

[20] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti del Congresso Internazionale dei Matematici*, pages 179–190, 1929.

[21] B. de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, pages 298–329, 1931.

[22] B. de Finetti. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, pages 1–68, 1937.

[23] B. de Finetti. Probabilism: A critical essay on the theory of probability and on the value of science. *Erkenntnis (translation of "Probabilismo. Saggio critico sulla teoria delle probabilita e sul valore della scienza", Biblioteca di Filosofia, Napoli, 1931)*, 31:169–223, 1989.

[24] A. de Moivre. *The doctrine of chances: or, a method of calculating the probability of events in play*. W. Pearson, London (reprinted 1967, New York, NY: Chelsea), 1718.

[25] L.E. Dubins and D.A. Freedman. Exchangeable processes need not be mixtures of independent, identically distributed random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 48:115–132, 1979.

[26] L. Euler. Pièce qui a remporté le prix de l'Académie Royale des Sciences en 1748, sur l'inégalités du movement de Saturn et de Jupiter. In *Leonhardi Euleri Opera Omnia*.

[27] J. Foster. *The divine lawmaker: Lectures on induction, laws of nature and the existence of god*. Clarendon Press, Oxford, 2004.

[28] W.J. Frable. Thin needle aspiration biopsy. In *Major Problems in Pathology*, volume 14. WB Saunders Co, Philadelphia, 1983.

[29] S. Garatti and M. C. Campi. Risk and complexity in scenario optimization. *Mathematical Programming – Series B*, 191:243–279, 2022.

[30] S. Garatti, M. C. Campi, and A. Carè. On a class of interval predictor models with universal reliability. *Automatica*, 110:1–9, 2019.

[31] S. Garatti and M.C. Campi. Non-convex scenario optimization. *Internal Report, University of Brescia*, 2023.

[32] S. Garatti and M.C. Campi. On conditional risk assessments in scenario optimization. *SIAM J. Control and Optim.*, 33(2):455–480, 2023.

[33] R.W.M. Giard and J. Hermans. The value of aspiration cytologic examination of the breast. *Cancer*, 69:2104–2110, 1992.

[34] N. Goodman. *Fact, Fiction, & Forecast.* Harvard University Press, 1955.

[35] A. Haar. Die Minkowskische geometrie und die annäherung an stetige funktionen. *Mathematische Annalen*, 78:294–311, 1918.

[36] I. Hacking. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 1975.

[37] H.L. Harter. The method of least squares and some alternatives – part iii. *International Statistical Reviews*, 43:1–44, 1975.

[38] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[39] J.C. Hull. *Options, futures and other derivatives*. Pearson/Prentice Hall (8th Edition), 2009.

[40] D. Hume. *An enquiry concerning human understanding*. Harvard Classics, Vol. 37, Part 3, Collier and Son, 1909-14 (originally published in 1748).

[41] D. Hume. *A treatise of human nature*. Oxford University Press, 1739.

[42] T.H. Huxley. *Agnosticism and Christianity and other essays*. Prometeus Books, 1992 (from "Agnosticism and Christianity", Collected Essays V, 1899).

[43] E.T. Jaynes. *Probability theory. The logic of science*. Cambridge University Press, 2003.

[44] H. Jeffreys. *Theory of probability*. Clarendon Press, Oxford, 1939.

[45] J.M. Keynes. *A Treatise on probability*. Macmillan, London, 1921.

[46] A.N. Kolmogorov. *Grundbegriffe der wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik und Ihrer Grenzgebiete, 1933 (translated in English as "Foundations of the theory of probability", Chelsea, 2nd edition, 1956).

[47] P.S. Laplace. *Mémoires de l'Académie Royale des Sciences*, 1783:17–46, 1783.

[48] P.S. Laplace. *Essai philosophique des probabilitès*. (translated version Philosophical Essay of Probabilities, Springer, 1999), 1814.

[49] V.L. Levin. Application of E. Helly's theorem to convex programming, problems of best approximation and related questions. *Mathematics of the USSR – Sbornik*, 8:235–247, 1969.

[50] P. Maher. The hole in the ground of induction. *Australasian Journal of Philosophy*, 74:423–432, 1996.

[51] A. De Morgan. *Formal logic, or, the calculus of inference, necessary and probable*. Taylor and Walton, London, 1847.

[52] H. Poincarè. *La Science et l'hypothèse*. Ernest Flammarion (English edition "Science and Hypothesis", Bloomsbury, 2017), 1904.

[53] K. Popper. *Conjectures and refutations: The growth of scientific knowledge*. Routledge & Kegan Paul, 1963.

[54] F.A. Ramponi and M.C. Campi. Expected shortfall: Heuristics and certificates. *European Journal of Operational Research*, 267:1003–1013, 2018.

[55] F.P. Ramsey. Truth and probability. In *"The foundations of mathematics and other logical essays", Ch. VII, 156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner and Co., 1931 (reprinted in H.E. Kyburg and H.E. Smokler (eds.), "Studies in subjective probability", 25-52, New York: Robert Krieger, 1980)*.

[56] K.P.S. Bhaskara Rao and M. Bhaskara Rao. *Theory of charges*. Academic Press, 1983.

[57] H. Reichenbach. *The theory of probability: An inquiry into the logical and mathematical foundations of the calculus of probability*. University of California Press, Berkeley (revised English edition of "Wahrscheinlichkeitslehre", published in German in 1935), 1949.

[58] R. T. Rockafellar and S. Uryasev. Optimization of conditional Value-at-Risk. *Journal of Risk*, 2:21–41, 2000.

[59] R. T. Rockafellar and S. Uryasev. Conditional Value-at-Risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

[60] B. Schölkopf and A.J. Smola. *Learning with kernels*. MIT press, 1998.

[61] A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1996.

[62] D. Steel. What if the principle of induction is normative? Formal learning theory and Hume's problem. *International Studies in the Philosophy of Science*, 24:171–185, 2010.

[63] D.C. Stove. *The rationality of induction*. Oxford University Press, 1986.

[64] W.N. Street, W.H. Wolberg, and O.L. Mangasarian. Nuclear feature extraction for breast ttumor diagnosis. In *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870. San Jose, CA, 1993.

[65] B. van Fraassen. *Laws and symmetry*. Clarendon Press, Oxford, 1989.

[66] J. Venn. *The logic of chance: An essay on the foundations and province of the theory of probability, with especial reference to its application to moral and social science*. Macmillan, London and Cambridge, 1866.

[67] R. von Mises. *Probability, statistics and truth*. Macmillan, New York (revised English edition of the second edition of "Wahrscheinlichkeit, Statistik und Wahrheit", published in German in 1928), 1957.

[68] D.C. Williams. *The ground of induction*. Harvard University Press, 1947.

Index

σ-additivity, 14

agnostic setup, 35, 37, 58, 77, 78
algebra, 14
  of events, 14
Artzner, Philippe, 50

Bayesian prior, 42
Bernoulli, Jacob, 23

Campi, Marco C., 48
Carnap, Rudolf, 29
Chebyshev layer, 45
Chebyshev, Pafnuty L., 46
classification
  binary, 52
complexity, 55
conditional assessment, 76, 78
Conditional Value at Risk – CVaR, 50, 71
conjectures and refutations
  theory of, 74
consistency, 11
  condition of, 49, 52
convex
  hull, 61, 63, 65
  optimization, 66, 67
  set, 61
corroboration rate, 75

de Finetti, Bruno, 22–24, 32, 74
de Moivre, Abraham, 23
de Morgan, Augustus, 22
degeneracy, 63
degree of belief, 22
distribution function
  cumulative, 67
dominance, 49

empirical bundle, 9
enumerative induction, 44, 61
Euler, Leonhard, 46
exchangeable observations, 73

Garatti, Simone, 48
generalization, 8, 10
Goodman, Nelson, 44
Guaranteed Error Machine – GEM, 52, 72

Haar, Alfréd, 46
Hacking, Ian, 42
Hoeffding, Wassily, 37
Hume, David, 26
Huxley, Thomas H., 35

Jaynes, Edwin T., 24
Jeffreys, Harold, 29

Keynes, John M., 23, 29
knowledge
  a priori, 36, 39, 60
  background, 13, 25
  conditional, 39, 42
  domain, 60
  from observations, 35
  scientific, 74
  uncertain, 13

update of, 15
Kolmogorov, Andrej N., 14

Laplace, Pierre S., 23, 46
law
of large numbers, 38
Levin, Vladimir Lvovich, 66
likelihood function, 80
list
exchangeable, 18
independent and identically distributed, 17

machine learning, 52
Maher, Patrick, 42
market
bear, 60
bull, 60
Markov kernel, 16
model
observation-driven, 7
of a population, 43

optimization
min-max, 9

Pascal, Blaise, 23
permutation invariance, 48
Poincaré, Henri, 23, 32
Popper, Karl, 74
portfolio, 9
loss, 9
principle
of indifference, 23, 30, 80
of optimality, 44
of parsimony, 5
of uniformity, 26
probabilities
set of, 30
probability, 13
conditional, 15, 40
elementary, 13
frequentist interpretation of, 28
image, 16
Kolmogorov's definition, 24
logical, 28
subjective, 22
total, 40
problem
fully supported, 66
with complexity d, 66
with limited complexity, 66
problem of induction, 26
Pólya's urn, 18
Pólya, George, 18

Ramsey, Frank P., 22
RandN, 32
random variable, 16
rate-of-return, 9, 28
Reichenbach, Hans, 28
represntation theorem, 74
risk, 56
upper and lower bound to, 64
upper bound to, 58
Rockafellar, R. Tyrrell, 50

shortfall threshold, 51
Steel, Daniel, 44
Stove, David, 42
support sub-sample, 49
irreducible, 63

test of hypothesis, 80
theorem
of Ionescu-Tulcea, 39
training set, 52

uncertainty
aleatoric, 22
epistemic, 22
Uryasev, Stanislav, 50

Venn, John, 28
von Leibniz, Gottfried W., 23
von Mises, Richard, 28

Williams, Donald C., 42