Fabio Cuzzolin

The geometry of uncertainty

The geometry of imprecise probabilities

October 7, 2020

Springer
Berlin Heidelberg New York
Barcelona Hong Kong
London Milan Paris
Tokyo

Preface

Uncertainty

Uncertainty is of paramount importance in artificial intelligence, applied science, and many other areas of human endeavour. Whilst each and every one of us possesses some intuitive grasp of what uncertainty is, providing a formal definition can prove elusive. Uncertainty can be understood as a lack of information about an issue of interest for a certain agent (e.g., a human decision maker or a machine), a condition of limited knowledge in which it is impossible to exactly describe the state of the world or its future evolution.

According to Dennis Lindley [1175]:

" *There are some things that you know to be true, and others that you know to be false; yet, despite this extensive knowledge that you have, there remain many things whose truth or falsity is not known to you. We say that you are uncertain about them. You are uncertain, to varying degrees, about everything in the future; much of the past is hidden from you; and there is a lot of the present about which you do not have full information. Uncertainty is everywhere and you cannot escape from it* ".

What is sometimes less clear to scientists themselves is the existence of a hiatus between two fundamentally distinct forms of uncertainty. The first level consists of somewhat '*predictable*' variations, which are typically encoded as probability distributions. For instance, if a person plays a fair roulette wheel they will not, by any means, know the outcome in advance, but they will nevertheless be able to predict the frequency with which each outcome manifests itself (1/36), at least in the long run. The second level is about '*unpredictable*' variations, which reflect a more fundamental uncertainty about the laws themselves which govern the outcome. Continuing with our example, suppose that the player is presented with ten different doors, which lead to rooms each containing a roulette wheel modelled by a different probability distribution. They will then be uncertain about the very game they are supposed to play. How will this affect their betting behaviour, for instance?

Lack of knowledge of the second kind is often called *Knightian* uncertainty [1007, 831], from the Chicago economist Frank Knight. He would famously distinguish 'risk' from 'uncertainty':

"*Uncertainty must be taken in a sense radically distinct from the familiar notion of risk, from which it has never been properly separated . . . The essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two*

is really present and operating . . . It will appear that a measurable uncertainty, or 'risk' proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all."
In Knight's terms, 'risk' is what people normally call *probability* or *chance*, while the term 'uncertainty' is reserved for second-order uncertainty. The latter has a measurable consequence on human behaviour: people are demonstrably averse to unpredictable variations (as highlighted by *Ellsberg's paradox* [569]).

This difference between predictable and unpredictable variation is one of the fundamental issues in the philosophy of probability, and is sometimes referred to as the distinction between *common cause* and *special cause* [1739]. Different interpretations of probability treat these two aspects of uncertainty in different ways, as debated by economists such as John Maynard Keynes [961] and G. L. S. Shackle.

Probability

Measure-theoretical probability, due to the Russian mathematician Andrey Kolmogorov [1030], is the mainstream mathematical theory of (first-order) uncertainty. In Kolmogorov's mathematical approach probability is simply an application of measure theory, and uncertainty is modelled using additive measures.

A number of authors, however, have argued that measure-theoretical probability theory is not quite up to the task when it comes to encoding second-order uncertainty. In particular, as we discuss in the Introduction, additive probability measures cannot properly model missing data or data that comes in the form of *sets*. Probability theory's frequentist interpretation is utterly incapable of modelling 'pure' data (without 'designing' the experiment which generates it). In a way, it cannot even properly model continuous data (owing to the fact that, under measure-theoretical probability, every point of a continuous domain has zero probability), and has to resort to the 'tail event' contraption to assess its own hypotheses. Scarce data can only be effectively modelled asymptotically.

Bayesian reasoning is also plagued by many serious limitations: (i) it just cannot model ignorance (absence of data); (ii) it cannot model pure data (without artificially introducing a prior, even when there is no justification for doing so); (iii) it cannot model 'uncertain' data, i.e., information not in the form of propositions of the kind 'A is true'; and (iv) again, it is able to model scarce data only asymptotically, thanks to the Bernstein–von Mises theorem [1841].

Beyond probability

Similar considerations have led a number of scientists to recognise the need for a coherent mathematical theory of uncertainty able to properly tackle all these issues. Both alternatives to and extensions of classical probability theory have been proposed, starting from de Finetti's pioneering work on subjective probability [403]. Formalisms include possibility-fuzzy set theory [2084, 533], probability intervals

[784], credal sets [1141, 1086], monotone capacities [1911], random sets [1344] and imprecise probability theory [1874]. New original foundations of subjective probability in behavioural terms [1877] or by means of game theory [1615] have been put forward. The following table presents a sketchy timeline of the various existing approaches to the mathematics of uncertainty.

Imprecise-probabilistic theories: a timeline

Approach	Proposer(s)	Seminal paper	Year
Interval probabilities	John Maynard Keynes	A treatise on probability	1921
Subjective probability	Bruno de Finetti	Sul significato soggettivo della probabilità	1931
Theory of previsions	Bruno de Finetti	La prévision: ses lois logiques, ses sources subjectives	1937
Theory of capacities	Gustave Choquet	Theory of capacities	1953
Fuzzy theory	Lotfi Zadeh, Dieter Klaua	Fuzzy sets	1965
Theory of evidence	Arthur Dempster, Glenn Shafer	Upper and lower probabilities induced by a multivalued mapping; A mathematical theory of evidence	1967, 1976
Fuzzy measures	Michio Sugeno	Theory of fuzzy integrals and its applications	1974
Credal sets	Isaac Levi	The enterprise of knowledge	1980
Possibility theory	Didier Dubois, Henri Prade	Théorie des possibilités	1985
Imprecise probability	Peter Walley	Statistical reasoning with imprecise probabilities	1991
Game-theoretical probability	Glenn Shafer, Vladimir Vovk	Probability and finance: It's only a game!	2001

Sometimes collectively referred to as *imprecise probabilities* (as most of them comprise classical probabilities as a special case), these theories in fact form, as we will see in more detail in Chapter 6, an entire hierarchy of encapsulated formalisms.

Belief functions

One of the most popular formalisms for a mathematics of uncertainty, the *theory of evidence* [1583] was introduced in the 1970s by Glenn Shafer as a way of representing epistemic knowledge, starting from a sequence of seminal papers [415, 417, 418] by Arthur Dempster [416]. In this formalism, the best representation of chance is a *belief function* rather than a traditional probability distribution. Belief functions assign probability values to *sets* of outcomes, rather than single events. In this sense, belief functions are closely related to *random sets* [1268, 1857, 826]. Important work on the mathematics of random set theory has been conducted in recent years by Ilya Molchanov [1302, 1304].

In its original formulation by Dempster and Shafer, the formalism provides a simple method for merging the evidence carried by a number of distinct sources (called *Dempster's rule of combination* [773]), with no need for any prior distributions [1949]. The existence of different levels of granularity in knowledge representation is formalised via the concept of a 'family of compatible frames'.

The reason for the wide interest the theory of belief functions has attracted over the years is that it addresses most of the issues probability theory has with the handling of second-order uncertainty. It starts from the assumption that observations are indeed set-valued and that evidence is, in general, in support of propositions rather than single outcomes. It can model ignorance by simply assigning mass to the whole sample space or 'frame of discernment'. It copes with missing data in the most natural of ways, and can coherently represent evidence on different but compatible sample spaces. It does not 'need' priors but can make good use of them whenever there is actual prior knowledge to exploit. As a direct generalisation of classical probability, the theory's rationale is relatively easier to grasp. Last but not least, the formalism does not require us to entirely abandon the notion of an event, as is the case for Walley's imprecise probability theory [1874]. In addition, the theory of evidence exhibits links to most other theories of uncertainty, as it includes fuzzy and possibility theory as a special case and it relates to the theory of credal sets and imprecise probability theory (as belief functions can be seen as a special case of convex sets of probability measures). Belief functions are infinitely-monotone capacities, and have natural interpretations in the framework of probabilistic logic, and modal logic in particular.

Since its inception, the formalism has expanded to address issues such as inference (how to map data to a belief function), conditioning, and the generalisation of the notion of entropy and of classical results from probability theory to the more complex case of belief functions. The question of what combination rule is most appropriate under what circumstances has been hotly debated, together with that of mitigating the computational complexity of working with sets of hypotheses. Graphical models, machine learning approaches and decision making frameworks based on belief theory have also been developed.

A number of questions still remain open, for instance on what is the correct epistemic interpretation of belief functions, whether we should actually manipulate intervals of belief functions rather than single quantities, and how to formulate an effective general theory for the case of continuous sample spaces.

Aim(s) of the book

The principal aim of this book is to introduce to the widest possible audience an original view of belief calculus and uncertainty theory which I first developed during my doctoral term in Padua. In this *geometric approach to uncertainty*, uncertainty measures can be seen as points of a suitably complex geometric space, and there manipulated (e.g. combined, conditioned and so on).

The idea first sprang to my mind just after I had been introduced to non-additive probabilities. Where did such objects live, I wondered, when compared to classical, additive probabilities defined on the same sample space? How is their greater complexity reflected in the geometry of their space? Is the latter an expression of the greater degree of freedom these more complex objects can provide?

For the reasons mentioned above, my attention was first drawn to belief functions and their combination rule which, from the point of view of an engineer, appeared to provide a possible principled solution to the sensor fusion problems one encounters in computer vision when making predictions or decisions based on multiple measurements or 'features'. Using the intuition gathered in the simplest case study of a binary domain, I then proceeded to describe the geometry of belief functions and their combination in fully general terms and to extend, in part, this geometric analysis to other classes of uncertainty measures.

This programme of work is still far from reaching its conclusion – nevertheless, I thought that the idea of consolidating my twenty-year work on the geometry of uncertainty in a monograph had some merit, especially in order to disseminate the notion and encourage a new generation of scientists to develop it further. This is the purpose of the core of the book, Parts II, III and IV.

In the years that it took for this project to materialise, I realised that the manuscript could serve the wider purpose of illustrating the rationale for moving away from probability theory to non-experts and interested practitioners, of which there are many. Incidentally, this forced me to reconsider from the very foundations the reasons for modelling uncertainty in a non-standard way. These reasons, as understood by myself, can be found in the Introduction, which is an extended version of the tutorial I gave on the topic at IJCAI 2016, the International Joint Conference on Artificial Intelligence, and the talk I was invited to give at Harvard University in the same year.

The apparent lack of a comprehensive treatise on belief calculus in its current, modern form (and, from a wider perspective, of uncertainty theory) motivated me to make use of this book to provide what turned out to be probably the most complete summary (to the best of my knowledge) of the theory of belief functions. The entire first part of the book is devoted to this purpose. Part I is not quite a 'manual' on belief calculus, with easy recipes the interested practitioner can just follow, but does strive to make a serious effort in that direction. Furthermore, the first part of the book concludes with what I believe to be the most complete compendium of the various approaches to uncertainty theory, with a specific focus on how do they relate to the theory of evidence. All major formalisms are described in quite some detail, but an effort was really made to cover, albeit briefly, all published approaches to a mathematics of uncertainty and variations thereof.

Finally, the last chapter of the book advances a tentative research agenda for the future of the field, inspired by my own reflections and ideas on this. As will become clearer in the remainder of this work, my intuition brings me to favour a random-set view of uncertainty theory, driven by an analysis of the actual issues with data that expose the limitations of probability theory. As a result, the research problems I propose tend to point in this direction. Importantly, I strongly believe that, to break the near-monopoly of probability theory in science, uncertainty theory needs to measure itself with the really challenging issues of our time (climate change, robust artificial intelligence), compete with mainstream approaches and demonstrate its superior expressive power on their own grounds.

Last but not least, the book provides, again to the best of my knowledge, the largest existing bibliography on belief and uncertainty theory.

Structure and topics

Accordingly, as explained, this book is divided into five Parts.

Part I, 'Theories of uncertainty', is a rather extensive recapitulation of the current state of the art in the mathematics of uncertainty, with a focus on belief theory. The Introduction provided in Chapter 1 motivates in more detail the need to go beyond classical probability in order to model realistic, second-order uncertainty, introduces the most significant approaches to the mathematics of uncertainty and presents the main principles of the theory of belief functions. Chapter 2 provides a succinct summary of the basic notions of the theory of belief functions as formulated by Shafer. Chapter 3 digs deeper by recalling the multiple semantics of belief functions, discussing the genesis of the approach and the subsequent debate, and illustrating the various original frameworks proposed by a number of authors which use belief theory as a basis, while developing it further in original ways. Chapter 4 can be thought of as a manual for the working scientist keen on applying belief theory. It illustrates in detail all the elements of the evidential reasoning chain, delving into all its aspects, including inference, conditioning and combination, efficient computation, decision making and continuous formulations. Notable advances in the mathematics of belief functions are also briefly described. Chapter 5 surveys the existing array of classification, clustering, regression and estimation tools based on belief function theory. Finally, Chapter 6 is designed to provide the reader with a bigger picture of the whole field of uncertainty theory, by reviewing all major formalisms (the most significant of which are arguably Walley's imprecise probability, the theory of capacities and fuzzy/possibility theory), with special attention paid to their relationship with belief and random set theory.

Part II, 'The geometry of uncertainty', is the core of this book, as it introduces the author's own geometric approach to uncertainty theory, starting with the geometry of belief functions. First, Chapter 7 studies the geometry of the space of belief functions, or *belief space*, both in terms of a simplex (a higher-dimensional triangle) and in terms of its recursive bundle structure. Chapter 8 extends the analysis to Dempster's rule of combination, introducing the notion of a conditional subspace and outlining a simple geometric construction for Dempster's sum. Chapter 9 delves into the combinatorial properties of plausibility and commonality functions, as equivalent representations of the evidence carried by a belief function. It shows that the corresponding spaces also behave like simplices, which are congruent to the belief space. The remaining Chapter 10 starts extending the applicability of the geometric approach to other uncertainty measures, focusing in particular on possibility measures (consonant belief functions) and the related notion of a consistent belief function.

Part III, 'Geometric interplays', is concerned with the interplay of uncertainty measures of different kinds, and the geometry of their relationship. Chapters 11 and 12 study the problem of transforming a belief function into a classical probability measure. In particular, Chapter 11 introduces the *affine family* of probability transformations, those which commute with affine combination in the belief space. Chapter 12 focuses instead on the *epistemic* family of transforms, namely 'relative belief' and 'relative plausibility', studies their dual properties with respect to Dempster's sum, and describes their geometry on both the probability simplex and the belief space. Chapter 13 extends the analysis to the consonant approximation problem, the problem of finding the possibility measure which best approximates a given belief function. In particular, approximations induced by classical Minkowski norms are derived, and compared with classical outer consonant approximations. Chapter 14 concludes Part III by describing Minkowski consistent approximations of belief functions in both the mass and the belief space representations.

Part IV, 'Geometric reasoning', examines the application of the geometric approach to the various elements of the reasoning chain illustrated in Chapter 4. Chapter 15 tackles the conditioning problem from a geometric point of view. Conditional belief functions are defined as those which minimise an appropriate distance between the original belief function and the 'conditioning simplex' associated with the conditioning event. Analytical expressions are derived for both the belief and the mass space representations, in the case of classical Minkowski distances. Chapter 16 provides a semantics for the main probability transforms in terms of credal sets, i.e., convex sets of probabilities. Based on this interpretation, decision-making apparatuses similar to Smets's transferable belief model are outlined.

Part V, 'The future of uncertainty', consisting of Chapter 17, concludes the book by outlining an agenda for the future of the discipline. A comprehensive statistical theory of random sets is proposed as the natural destination of this evolving field. A number of open challenges in the current formulation of belief theory are highlighted, and a research programme for the completion of our geometric approach to uncertainty is proposed. Finally, very high-impact applications in fields such as climate change, rare event estimation, machine learning and statistical learning theory are singled out as potential triggers of a much larger diffusion of these techniques and of uncertainty theory in general.

Acknowledgements

This book would not have come to existence had I not, during my doctoral term at the University of Padua, been advised by my then supervisor Ruggero Frezza to attend a seminar by Claudio Sossai at Consiglio Nazionale delle Ricerche (CNR). At the time, as every PhD student in their first year, I was struggling to identify a topic of research to focus on, and felt quite overwhelmed by how much I did not know about almost everything. After Claudio's seminar I was completely taken by the idea of non-additive probability, and ended up dedicating my entire doctorate to the study of the mathematics of these beautiful objects. Incidentally, I need to thank Nicola Zingirian and Riccardo Bernardini for enduring my ramblings during our coffee breaks back then.

I would like to acknowledge the encouragement I received from Glenn Shafer over the years, which has greatly helped me in my determination to contribute to belief theory in terms of both understanding and impact. I would also like to thank Art Dempster for concretely supporting me in my career and for the insights he shared with me in our epistolary exchanges (not to mention the amazing hospitality displayed during my visit to Harvard in 2016).

Running the risk of forgetting somebody important, I would like to thank Teddy Seidenfeld, Gert de Cooman, Matthias Troffaes, Robin Gong, Xiao-Li Meng, Frank Coolen, Thierry Denoeux, Paul Snow, Jonathan Lawry, Arnaud Martin, Johan Schubert, Anne-Laure Jousselme, Sebastien Destercke, Milan Daniel, Jim Hall, Alberto Bernardini, Thomas Burger and Alessandro Antonucci for the fascinating conversations that all deeply influenced my work. I am also thankful to Ilya Molchanov for his help with better understanding the links between belief function theory and random set theory.

Finally, I am grateful to my Springer editor Ronan Nugent for his patience and support throughout all those years that passed from when we started this project to the moment we could actually cheer its successful outcome.

Oxford, United Kingdom *Fabio Cuzzolin*

September 2020

Table of Contents

Part I Theories of uncertainty

Part II The geometry of uncertainty

Part III Geometric interplays

Part IV Geometric reasoning

Part V The future of uncertainty

1
Introduction

1.1 Mathematical probability

The mainstream mathematical theory of uncertainty is measure-theoretical probability, and is mainly due to the Russian mathematician Andrey Kolmogorov [1030]. As most readers will know, in Kolmogorov's mathematical approach[1] probability is simply an application of *measure theory* [783], the theory of assigning numbers to sets. In particular, Kolmogorov's probability measures are *additive* measures, i.e., the real value assigned to a set of outcomes is the sum of the values assigned to its constituent elements. The collection Ω of possible outcomes (of a random experiment or of a decision problem) is called the *sample space*, or universe of discourse. Any (measurable) subset A of the universe Ω is called an *event*, and is assigned a real number between 0 and 1.

Formally [1030], let Ω be the sample space, and let 2^{Ω} represent its power set $2^{\Omega} \doteq \{A \subset \Omega\}$. The power set is also sometimes denoted by $\mathcal{P}(\Theta)$.

Definition 1. *A collection $\mathcal{F}$ of subsets of the sample space, $\mathcal{F} \subset 2^{\Omega}$, is called a* σ-algebra *or* σ-field *if it satisfies the following three properties:*

- *$\mathcal{F}$ is non-empty: there is at least one $A \subset \Omega$ in $\mathcal{F}$;*
- *$\mathcal{F}$ is closed under complementation: if A is in $\mathcal{F}$, then so is its complement, $\overline{A} = \{\omega \in \Omega, \omega \notin A\} \in \mathcal{F}$;*
- *$\mathcal{F}$ is closed under countable union: if $A_1, A_2, A_3, \cdots$ are in $\mathcal{F}$, then so is $A = A_1 \cup A_2 \cup A_3 \cup \cdots$,*

[1] A recent study of the origins of Kolmogorov's work has been done by Shafer and Vovk: `http://www.probabilityandfinance.com/articles/04.pdf`.

where $\cup$ *denotes the usual set-theoretical union.*

Any subset of Ω which belongs to such a σ-algebra is said to be *measurable*. From the above properties, it follows that any σ-algebra $\mathcal{F}$ is closed under *countable intersection* as well by De Morgan's laws: $\overline{A \cup B} = \overline{A} \cap \overline{B}$, $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Definition 2. *A* probability measure *over a σ-algebra $\mathcal{F} \subset 2^{\Omega}$, associated with a sample space Ω, is a function $P : \mathcal{F} \to [0,1]$ such that:*

- $P(\emptyset) = 0$*;*
- $P(\Omega) = 1$*;*
- *if* $A \cap B = \emptyset,\ A, B \in \mathcal{F}$ *then* $P(A \cup B) = P(A) + P(B)$ *(*additivity*).*

A simple example of a probability measure associated with a spinning wheel is shown in Fig. 1.1.

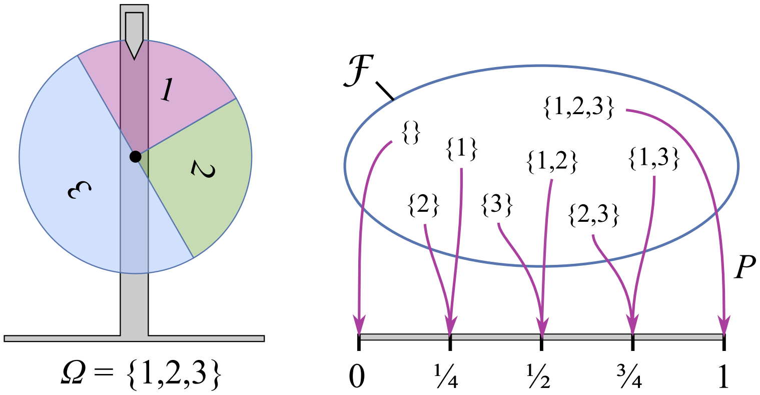

Fig. 1.1: A spinning wheel is a physical mechanism whose outcomes are associated with a (discrete) probability measure (adapted from original work by Ziggystar, `https://commons.wikimedia.org/wiki/File:Probability-measure.svg`).

A sample space Ω together with a σ-algebra $\mathcal{F}$ of its subsets and a probability measure P on $\mathcal{F}$ forms a *probability space*, namely the triplet $(\Omega, \mathcal{F}, P)$. Based on the notion of a probability space, one can define that of a *random variable*. A random variable is a quantity whose value is subject to random variations, i.e., to 'chance' (although, as we know, what chance is is itself subject to debate). Mathematically, it is a function X from a sample space Ω (endowed with a probability space) to a measurable space E (usually the real line $\mathbb{R}$)[2]. Figure 1.4 (left) illustrates the random variable associated with a die.

[2] However, the notion of a random variable can be generalised to include mappings from a sample space to a more structured domain, such as an algebraic structure. These functions are called *random elements* [643].

To be a random variable, a function $X : \Omega \to \mathbb{R}$ must be *measurable*: each measurable set in E must have a pre-image $X^{-1}(E)$ which belongs to the σ-algebra $\mathcal{F}$, and therefore can be assigned a probability value. In this way, a random variable becomes a means to assign probability values to sets of real numbers.

1.2 Interpretations of probability

1.2.1 Does probability exist at all?

When one thinks of classical examples of probability distributions (e.g. a spinning wheel, a roulette wheel or a rolling die), the suspicion that 'probability' is simply a fig leaf for our ignorance and lack of understanding of nature phenomena arises.

Assuming a view of the physical world that follows the laws of classical Newtonian mechanics, it is theoretically conceivable that perfect knowledge of the initial conditions of say, a roulette wheel, and of the impulse applied to it by the croupier would allow the player to know exactly what number would come out. In other words, with sufficient information, any phenomenon would be predictable in a completely deterministic way. This is a position supported by Einstein himself, as he was famously quoted as saying that 'God does not play dice with the universe'. In Doc Smith's Lensman series [1736], the ancient race of the Arisians have such mental powers that they compete with each other over foreseeing events far away in the future to the tiniest detail.

A first objection to this argument is that 'infinite accuracy' is an abstraction, and any actual measurements are bound to be affected by a degree of imprecision. As soon as initial states are not precisely known, the nonlinear nature of most phenomena inexcapably generates a chaotic behaviour that effectively prevents any accurate prediction of future events. More profoundly, the principles of quantum mechanics seem to suggest that probability is not just a figment of our mathematical imagination, or a representation of our ignorance: the workings of the physical world seem to be inherently probabilistic [119]. However, the question arises of why the finest structure of the physical world should be described by *additive* measures, rather than more general ones (or *capacities*: see Chapter 6).

Finally, as soon as we introduce the human element into the picture, any hope of being able to predict the future deterministically disappears. One may say that this is just another manifestation of our inability to understanding the internal workings of a system as complex as a human mind. Fair enough. Nevertheless, we still need to be able to make useful predictions about human behaviour, and 'probability' in a wide sense, is a useful means to that end.

1.2.2 Competing interpretations

Even assuming that (some form of mathematical) probability is inherent in the physical world, people cannot agree on what it is. Quoting Savage [45]:

> "*It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis*".

As a result, probability has multiple competing interpretations: (1) as an objective description of frequencies of events (meaning 'things that happen') at a certain persistent rate, or 'relative frequency' – this is the so-called *frequentist* interpretation, mainly due to Fisher and Pearson; (2) as a degree of belief in events (interpreted as statements/propositions on the state of the world), regardless of any random process – the *Bayesian* or *evidential* interpretation, first proposed by de Finetti and Savage; and (3) as the propensity of an agent to act (or gamble or decide) if the event happens – an approach called *behavioural* probability [1874].

Note that neither the frequentist nor the Bayesian approach is in contradiction with the classical mathematical definition of probability due to Kolmogorov: as we will see in this book, however, other approaches to the mathematics of uncertainty do require us to introduce different classes of mathematical objects.

1.2.3 Frequentist probability

In the frequentist interpretation, the (aleatory) probability of an event is its relative frequency in time. When one is tossing a fair coin, for instance, frequentists say that the probability of getting a head is 1/2, not because there are two equally likely outcomes (due to the structure of the object being tossed), but because repeated series of large numbers of trials (a *random experiment*) demonstrate that the empirical frequency converges to the limit 1/2 as the number of trials goes to infinity.

Clearly, it is impossible to actually complete the infinite series of repetitions which constitutes a random experiment. Guidelines on the design of 'practical' random experiments are nevertheless provided, via either *statistical hypothesis testing* or *confidence interval analysis*.

Statistical hypothesis testing A *statistical hypothesis* is a conjecture on the state of the world which is testable on the basis of observing a phenomenon modelled by a set of random variables.

In hypothesis testing, a dataset obtained by sampling is compared with data generated by an idealised model. A hypothesis about the statistical relationship between the two sets of data is proposed, and compared with an idealised 'null' hypothesis which rejects any relationship between the two datasets. The comparison is considered statistically significant if the relationship between the datasets would be an unlikely realisation of the null hypothesis according to a threshold probability, called the *significance level*. Statistical hypothesis testing is thus a form of confirmatory data analysis.

The steps to be followed in hypothesis testing are:[3]

[3] https://en.wikipedia.org/wiki/Statistical_hypothesis_testing.

1. State the null hypothesis H_0 and the alternative hypothesis H_1.
2. State the statistical assumptions being made about the sample, e.g. assumptions about the statistical independence or the distributions of the observations.
3. State the relevant *test statistic* T (i.e., a quantity derived from the sample).
4. Derive from the assumptions the distribution of the test statistic under the null hypothesis.
5. Set a significance level (α), i.e., a probability threshold below which the null hypothesis is rejected.
6. Compute from the observations the observed value t_{obs} of the test statistic T.
7. Calculate the *p-value*, the probability (under the null hypothesis) of sampling a test statistic 'at least as extreme' as the observed value.
8. Reject the null hypothesis, in favour of the alternative one, if and only if the p-value is less than the significance level threshold.

In hypothesis testing, false positives (i.e., rejecting a valid hypothesis) are called 'type I' errors; false negatives (not rejecting a false hypothesis) are called 'type II' errors. Note that if the p-value is above α, the result of the test is inconclusive: the evidence is insufficient to support a conclusion.

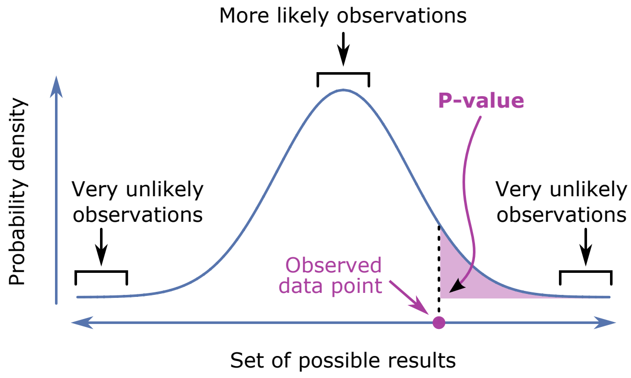

Fig. 1.2: Notion of p-value (adapted from `https://upload.wikimedia.org/wikipedia/commons/3/3a/P-value_in_statistical_significance_testing.svg`).

P-values The notion of a p-value is crucial in hypothesis testing. It is the probability, under the assumption of hypothesis H, of obtaining a result equal to or more extreme than what was actually observed, namely $P(X \geq x|H)$, where x is the observed value (see Fig. 1.2).

The reason for not simply considering $P(X = x|H)$ when assessing the null hypothesis is that, for any continuous random variable, such a conditional probability is equal to zero. As a result we need to consider, depending on the situation, a right-tail event $p = \mathbb{P}(X \geq x|H)$, a left-tail event $p = \mathbb{P}(X \leq x|H)$, or a double-tailed event: the 'smaller' of $\{X \leq x\}$ and $\{X \geq x\}$.

Note that the p-value is not the probability that the null hypothesis is true or the probability that the alternative hypothesis is false: frequentist statistics does not and cannot (by design) attach probabilities to hypotheses.

Maximum likelihood estimation A popular tool for estimating the parameters of a probability distribution which best fits a given set of observations is *maximum likelihood estimation* (MLE). The term *likelihood* was coined by Ronald Fisher in 1922 [620]. He argued against the use of 'inverse' (Bayesian) probability as a basis for statistical inferences, proposing instead inferences based on likelihood functions.

Indeed, MLE is based on the *likelihood principle*: all of the evidence in a sample relevant to model parameters is contained in the likelihood function. Some widely used statistical methods, for example many significance tests, are not consistent with the likelihood principle. The validity of such an assumption is still debated.

Definition 3. *Given a* parametric model $\{f(.|\theta), \theta \in \Theta\}$, *a family of conditional probability distributions of the data given a (vector) parameter* θ, *the maximum likelihood estimate of* θ *is defined as*

$$\hat{\theta}_{\mathrm{MLE}} \subseteq \left\{ \arg\max_{\theta \in \Theta} \mathcal{L}(\theta\,;\, x_1, \ldots, x_n) \right\},$$

where the likelihood of the parameter given the observed data $x_1, \ldots, x_n$ *is*

$$\mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta).$$

Maximum likelihood estimators have no optimal properties for finite samples: they do have, however, good asymptotic properties:

- *consistency*: the sequence of MLEs converges in probability, for a sufficiently large number of observations, to the (actual) value being estimated;
- *asymptotic normality*: as the sample size increases, the distribution of the MLE tends to a Gaussian distribution with mean equal to the true parameter (under a number of conditions[4]);
- *efficiency*: MLE achieves the *Cramer–Rao lower bound* [1841] when the sample size tends to infinity, i.e., no consistent estimator has a lower asymptotic mean squared error than MLE.

1.2.4 Propensity

The propensity theory of probability [1415], in opposition, thinks of probability as a *physical* propensity or tendency of a physical system to deliver a certain outcome. In a way, propensity is an attempt to explain why the relative frequencies of a random experiment turn out to be what they are. The law of large numbers is interpreted

[4] http://sites.stat.psu.edu/~dhunter/asymp/fall2003/lectures/pages76to79.pdf.

as evidence of the existence of invariant single-run probabilities (as opposed to the relative frequencies of the frequentist interpretation), which do emerge in quantum mechanics, for instance, and to which relative frequencies tend at infinity.

What propensity exactly means remains an open issue. Popper, for instance, has proposed a theory of propensity, which is, however, plagued by the use of relative frequencies in its own definition [1439].

1.2.5 Subjective and Bayesian probability

In *epistemic* or *subjective* probability, probabilities are degrees of belief assigned to events by an individual assessing the state of the world, whereas in frequentist inference a hypothesis is typically tested without being assigned a probability.

The most popular theory of subjective probability is perhaps the *Bayesian* framework [419], due to the English clergyman Thomas Bayes (1702–1761). There, all degrees of belief are encoded by additive mathematical probabilities (in Kolmogorov's sense). It is a special case of *evidential* probability, in which some *prior* probability is updated to a *posterior* probability in the light of new evidence (data). In the Bayesian framework, *Bayes' rule* is used sequentially to compute a posterior distribution when more data become available, namely whenever we learn that a certain proposition A is true:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{1.1}$$

Considered as an operator, Bayes' rule is inextricably related to the notion of *conditional* probability $P(B|A)$ [1143].

Bayes proved a special case of what is now called Bayes' theorem (1.1) in a paper entitled 'An essay towards solving a problem in the doctrine of chances'. Pierre-Simon Laplace (1749–1827) later introduced a general version of the theorem. Jeffreys' 'Theory of Probability' [894] (1939) played an important role in the revival of the Bayesian view of probability, followed by publications by Abraham Wald [1870] (1950) and Leonard J. Savage [1537] (1954).

The statistician Bruno de Finetti produced a justification for the Bayesian framework based on the notion of a *Dutch book* [404]. A Dutch book is made when a clever gambler places a set of bets that guarantee a profit, no matter what the outcome of the bets themselves. If a bookmaker follows the rules of Bayesian calculus, de Finetti argued, a Dutch book cannot be made. It follows that subjective beliefs must follow the laws of (Kolmogorov's) probability if they are to be coherent.

However, Dutch book arguments leave open the possibility that non-Bayesian updating rules could avoid Dutch books – one of the purposes of this book is indeed to show that this is the case. Justification by axiomatisation has been tried, but with no great success. Moreover, evidence casts doubt on the assumption that humans maintain coherent beliefs or behave rationally at all. Daniel Kahneman[5] won a Nobel Prize for supporting the exact opposite thesis, in collaboration with Amos

[5] https://en.wikipedia.org/wiki/Daniel_Kahneman.

Tversky. People consistently pursue courses of action which are bound to damage them, as they do not understand the full consequences of their actions.

For all its faults (as we will discuss later), the Bayesian framework is rather intuitive and easy to use, and capable of providing a number of 'off-the-shelf' tools to make inferences or compute estimates from time series.

Bayesian inference In Bayesian inference, the prior distribution is the distribution of the parameter(s) before any data are observed, i.e. $p(\theta|\alpha)$, a function of a vector of hyperparameters α. The likelihood is the distribution of the observed data $\mathbf{X} = \{x_1, \cdots, x_n\}$ conditional on its parameters, i.e., $p(\mathbf{X}|\theta)$. The distribution of the observed data marginalised over the parameter(s) is termed the *marginal likelihood* or *evidence*, namely

$$p(\mathbf{X}|\alpha) = \int_\theta p(\mathbf{X}|\theta)p(\theta|\alpha)\,\mathrm{d}\theta.$$

The *posterior distribution* is then the distribution of the parameter(s) after taking into account the observed data, as determined by Bayes' rule (1.1):

$$p(\theta|\mathbf{X},\alpha) = \frac{p(\mathbf{X}|\theta)p(\theta|\alpha)}{p(\mathbf{X}|\alpha)} \propto p(\mathbf{X}|\theta)p(\theta|\alpha). \tag{1.2}$$

The *posterior predictive distribution* is the distribution of a new data point x', marginalised over the posterior:

$$p(x'|\mathbf{X},\alpha) = \int_\theta p(x'|\theta)p(\theta|\mathbf{X},\alpha)\,\mathrm{d}\theta,$$

amounting to a distribution over possible new data values. The *prior predictive distribution*, instead, is the distribution of a new data point marginalised over the prior:

$$p(x'|\alpha) = \int_\theta p(x'|\theta)p(\theta|\alpha)\,\mathrm{d}\theta.$$

By comparison, prediction in frequentist statistics often involves finding an optimum point estimate of the parameter(s) (e.g., by maximum likelihood), not accounting for any uncertainty in the value of the parameter. In opposition, (1.2) provides as output an entire probability distribution over the parameter space.

Maximum a posteriori estimation *Maximum a posteriori (MAP) estimation* estimates a single value θ for the parameter as the mode of the posterior distribution (1.2):

$$\hat{\theta}_{\mathrm{MAP}}(x) \doteq \arg\max_\theta \frac{p(x|\theta)\,p(\theta)}{\displaystyle\int_\vartheta p(x|\vartheta)\,p(\vartheta)\,\mathrm{d}\vartheta} = \arg\max_\theta p(x|\theta)\,p(\theta).$$

MAP estimation is not very representative of Bayesian methods, as the latter are characterised by the use of distributions over parameters to draw inferences.

1.2.6 Bayesian versus frequentist inference

Summarising, in frequentist inference unknown parameters are often, but not always, treated as having fixed but unknown values that are not capable of being treated as random variates. Bayesian inference, instead, allows probabilities to be associated with unknown parameters. The frequentist approach does not depend on a subjective prior that may vary from one investigator to another. However, Bayesian inference (e.g. Bayes' rule) can be used by frequentists.[6]

Lindley's paradox is a counter-intuitive situation which occurs when the Bayesian and frequentist approaches to a hypothesis-testing problem give different results for certain choices of the prior distribution.

More specifically, Lindley's paradox[7] occurs when:

- the result x is 'significant' by a frequentist test of H_0, indicating sufficient evidence to reject H_0 say, at the 5% level, while at the same time
- the posterior probability of H_0 given x is high, indicating strong evidence that H_0 is in better agreement with x than H_1.

This can happen when H_0 is very specific and H_1 less so, and the prior distribution does not strongly favour one or the other.

It is not really a paradox, but merely a consequence of the fact that the two approaches answer fundamentally different questions. The outcome of Bayesian inference is typically a probability distribution on the parameters, given the results of the experiment. The result of frequentist inference is either a 'true or false' (binary) conclusion from a significance test, or a conclusion in the form that a given confidence interval, derived from the sample, covers the true value.

Glenn Shafer commented on the topic in [1590].

1.3 Beyond probability

A long series of students have argued that a number of serious issues arise whenever uncertainty is handled via Kolmogorov's measure-theoretical probability theory. On top of that, one can argue that something is wrong with both mainstream approaches to probability interpretation. Before we move on to introduce the mathematics of belief functions and other alternative theories of uncertainty, we think it best to briefly summarise our own take on these issues here.

1.3.1 Something is wrong with probability

Flaws of the frequentistic setting The setting of frequentist hypothesis testing is rather arguable. First of all, its scope is quite narrow: rejecting or not rejecting

[6] See for instance `www.stat.ufl.edu/~casella/Talks/BayesRefresher.pdf`.

[7] See `onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a15076/pdf`.

a hypothesis (although confidence intervals can also be provided). The criterion according to which this decision is made is arbitrary: who decides what an 'extreme' realisation is? In other words, who decides what is the right choice of the value of α? What is the deal with 'magical' numbers such as 0.05 and 0.01? In fact, the whole 'tail event' idea derives from the fact that, under measure theory, the conditional probability (p-value) of a point outcome is zero – clearly, the framework seems to be trying to patch up what is instead a fundamental problem with the way probability is mathematically defined. Last but not least, hypothesis testing cannot cope with pure data, without making additional assumptions about the process (experiment) which generates them.

The issues with Bayesian reasoning Bayesian reasoning is also flawed in a number of ways. It is extremely bad at representing ignorance: Jeffreys' uninformative priors [895] (e.g., in finite settings, uniform probability distributions over the set of outcomes), the common way of handling ignorance in a Bayesian setting, lead to different results for different reparameterisations of the universe of discourse. Bayes' rule assumes the new evidence comes in the form of certainty, 'A is true': in the real world, this is not often the case. As pointed out by Klir, a precise probabilistic model defined only on some class of events determines only interval probabilities for events outside that class (as we will discuss in Section 3.1.3).

Finally, model selection is troublesome in Bayesian statistics: whilst one is forced by the mathematical formalism to pick a prior distribution, there is no clear-cut criterion for how to actually do that.

In the author's view, this is the result of a fundamental confusion between the original Bayesian description of a person's subjective system of beliefs and the way it is updated, and the 'objectivist' view of Bayesian reasoning as a rigorous procedure for updating probabilities when presented with new information.

1.3.2 Pure data: Beware of the prior

Indeed, Bayesian reasoning requires modelling the data *and* a prior. Human beings do have 'priors', which is just a word for denoting what they have learned (or think they have learned) about the world during their existence. In particular, they have well-sedimented beliefs about the likelihood of various (if not all) events. There is no need to 'pick' a prior, for prior (accumulated) knowledge is indeed there. As soon as we idealise this mechanism to, say, allow a machine to reason in this way, we find ourselves forced to 'pick' a prior for an entity (an algorithm) which does not have any past experience, and has not sedimented any beliefs as a result. Nevertheless, Bayesians content themselves by claiming that all will be fine in the end, as, asymptotically, the choice of the prior does not matter, as proven by the Bernstein–von Mises theorem [1841].

1.3.3 Pure data: Designing the universe?

The frequentist approach, on its side, is inherently unable to describe pure data without having to make additional assumptions about the data-generating process.

Unfortunately, in nature one cannot 'design' the process which produces the data: data simply come our way. In the frequentist terminology, in most applications we cannot set the 'stopping rules' (think of driverless cars, for instance). Once again, the frequentist setting brings to the mind the image of a nineteenth-century scientist 'analysing' (from the Greek elements *ana* and *lysis*, breaking up) a specific aspect of the world within the cosy confines of their own laboratory.

Even more strikingly, it is well known that the same data can lead to opposite conclusions when analysed in a frequentist way. The reason is that different random experiments can lead to the same data, whereas the parametric model employed (the family of probability distributions $f(.|\theta)$ which is assumed to produce the data) is linked to a specific experiment.[8]

Apparently, however, frequentists are just fine with this [2131].

1.3.4 No data: Modelling ignorance

The modelling of ignorance (absence of data) is a major weakness of Bayesian reasoning. The typical solution is to pick a so-called 'uninformative' prior distribution, in particular *Jeffreys' prior*, the Gramian of the Fisher information matrix $\mathcal{I}$ [895]:

$$p(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}, \quad \mathcal{I}(\theta) \doteq E\left[\left(\frac{\partial}{\partial\theta} \log f(\mathbb{X}|\theta) \right)^2 \middle| \theta \right]. \tag{1.3}$$

Unfortunately, Jeffreys' priors can be improper (unnormalised). Most importantly, they violate the strong version of the likelihood principle: when using Jeffreys' prior, inferences about a parameter θ depend not just on the probability of the observed data as a function of θ, *but also on the universe Ω* of all possible experimental outcomes. The reason is that the Fisher information matrix $\mathcal{I}(\theta)$ is computed from an expectation (see (1.3)) over the chosen universe of discourse.

This flaw was pointed out by Glenn Shafer in his landmark book [1583], where he noted how the Bayesian formalism cannot handle multiple hypothesis spaces ('families of compatible frames', in Shafer's terminology: see Section 2.5.2) in a consistent way.

In Bayesian statistics, to be fair, one can prove that the asymptotic distribution of the posterior mode depends only on the Fisher information and not on the prior – the so-called *Bernstein–von Mises theorem*. The only issue is that the amount of information supplied must be large enough. The result is also subject to the caveat [644] that the Bernstein–von Mises theorem does not hold almost surely if the random variable considered has a countably infinite probability space.
As A. W. F. Edwards put it [564]:

"*It is sometimes said, in defence of the Bayesian concept, that the choice of prior distribution is unimportant in practice, because it hardly influences the posterior distribution at all when there are moderate amounts of data. The less said*

[8] http://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-\and-statistics-spring-2014/readings/MIT18_05S14_Reading20.pdf

about this 'defence' the better."

In actual fact, 'uninformative' priors can be dangerous, i.e., they may bias the reasoning process so badly that it can recover only asymptotically.[9]

As we will see in this book, instead, reasoning with belief functions does not require any prior. Belief functions encoding the available evidence are simply combined as they are, whereas ignorance is naturally represented by the 'vacuous' belief function, which assigns a mass equal to 1 to the whole hypothesis space.

1.3.5 Set-valued observations: The cloaked die

A die (Fig. 1.3) provides a simple example of a (discrete) random variable. Its probability space is defined on the sample space $\Omega = \{\text{face}1, \text{face } 2, \ldots, \text{face } 6\}$, whose elements are mapped to the real numbers $1, 2, \cdots, 6$, respectively (no need to consider measurability here).

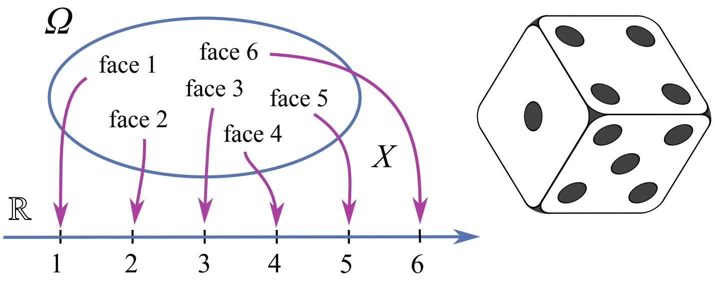

Fig. 1.3: The random variable X associated with a die.

Now, imagine that faces 1 and 2 are cloaked, and we roll the die. How do we model this new experiment, mathematically? Actually, the probability space has not changed (as the physical die has not been altered, its faces still have the same probabilities). What has changes is the mapping: since we cannot observe the outcome when a cloaked face is shown (we assume that only the top face is observable), both face 1 and face 2 (as elements of Ω) are mapped to the set of possible values $\{1, 2\}$ on the real line $\mathbb{R}$ (Fig. 1.4). Mathematically, this is called a *random set* [1268, 950, 1344, 1304], i.e., a set-valued random variable.

A more realistic scenario is that in which we roll, say, four dice in such a way that for some the top face is occluded, but some of the side faces are still visible, providing information about the outcome. For instance, I might be able to see the top face of the Red die as ⚄, the Green die as ⚂ and the Purple die as ⚃ but, say, not the outcome of the Blue die. Still, if I happen to observe the side faces ⚂ and ⚀ of Blue, I can deduce that the outcome of Blue is in the set $\{2, 4, 5, 6\}$.

[9] `http://andrewgelman.com/2013/11/21/hidden-dangers-noninformative-priors/`.

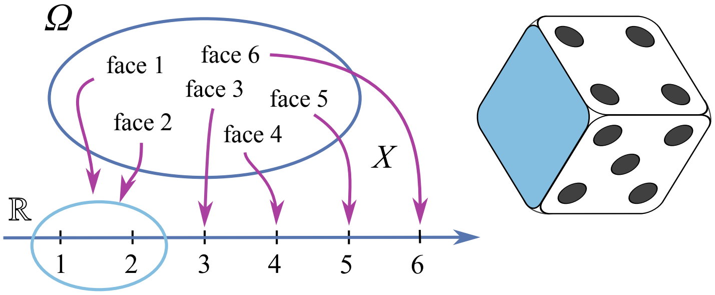

Fig. 1.4: The random set (set-valued random variable) associated with the cloaked die in which faces 1 and 2 are not visible.

This is just an example of a very common situation called *missing data*: for part of the sample that I need to observe in order to make my inference, the data are partly or totally missing. Missing data appear (or disappear?) everywhere in science and engineering. In computer vision, for instance, this phenomenon is typically called ‘occlusion’ and is one of the main nuisance factors in estimation.

The bottom line is, whenever data are missing, observations are inherently set-valued. Mathematically, we are sampling not a (scalar) random variable but a set-valued random variable – a random set. My outcomes are sets? My probability distribution has to be defined on sets.

In opposition, traditional statistical approaches deal with missing data either by deletion (discarding any case that has a missing value, which may introduce bias or affect the representativeness of the results); by single imputation (replacing a missing value with another one, e.g. from a randomly selected similar record in the same dataset, with the mean of that variable for all other cases, or by using a stochastic regression model); or multiple imputation (averaging the outcomes across multiple imputed datasets using, for instance, stochastic regression). Multiple imputation involves drawing values of the parameters from a posterior distribution, therefore simulating both the process generating the data and the uncertainty associated with the parameters of the probability distribution of the data.

When using random sets, there is no need for imputation or deletion whatsoever. All observations are set-valued: some of them just happen to be pointwise. Indeed, when part of the data used to estimate a probability distribution is missing, it has been shown that what we obtain instead is a convex set of probabilities or *credal set* [1141] (see Section 3.1.4), of the type associated with a *belief function* [401].

1.3.6 Propositional data

Just as measurements are naturally set-valued, in various scenarios evidence is directly supportive of propositions. Consider the following classical example [1607].

Suppose there is a murder, and three people are on trial for it: Peter, John and Mary. Our hypothesis space is therefore $\Theta = \{\text{Peter}, \text{John}, \text{Mary}\}$. There is a witness: he testifies that the person he saw was a man. This amounts to supporting the proposition $A = \{\text{Peter}, \text{John}\} \subset \Theta$. However, should we take this testimony at face value? In fact, the witness was tested and the machine reported an 80% chance that he was drunk when he reported the crime. As a result, we should partly support the (vacuous) hypothesis that any one among Peter, John and Mary could be the murderer. It seems sensible to assign 80% chance to proposition A, and 20% chance to proposition Θ (compare Chapter 2, Fig. 2.1).

This example tells us that, even when the evidence (our data) supports whole *propositions*, Kolmogorov's additive probability theory forces us to specify support for *individual outcomes*. This is unreasonable – an artificial constraint due to a mathematical model that is not general enough. In the example, we have no elements to assign this 80% probability to either Peter or John, nor information on how to distribute it among them. The cause is the additivity constraint that probability measures are subject to.

Kolmogorov's probability measures, however, are not the only or the most general type of measure available for sets. Under a minimal requirement of *monotonicity*, any measure can potentially be suitable for describing probabilities of events: the resulting mathematical objects are called *capacities* (see Fig. 1.5). We will study capacities in more detail in Chapter 6. For the moment, it suffices to note that random sets are capacities, those for which the numbers assigned to events are given by a probability distribution. Considered as capacities (and random sets in particular), belief functions therefore naturally allow us to assign mass directly to propositions.

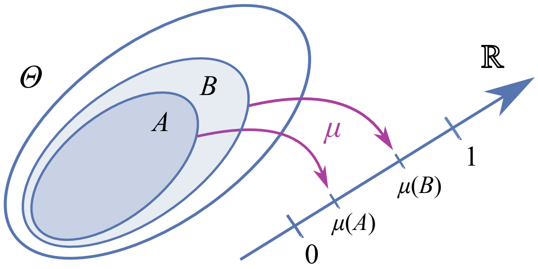

Fig. 1.5: A capacity μ is a mapping from 2^Θ to $[0, 1]$, such that if $A \subset B$, then $\mu(A) \leq \mu(B)$.

1.3.7 Scarce data: Beware the size of the sample

The current debate on the likelihood of biological life in the universe is an extreme example of inference from very scarce data. How likely is it for a planet to give birth to life forms? Modern analysis of planetary habitability is largely an extrapolation of conditions on Earth and the characteristics of the Solar System: a weak form of the old anthropic principle, so to speak. What people seem to do is model perfectly the (presumed) causes of the emergence of life on Earth: the planet needs to circle a G-class star, in the right galactic neighbourhood, it needs to be in a certain habitable zone around a star, have a large moon to deflect hazardous impact events, . . . The question arises: how much can one learn from a single example? More, how much can one be sure about what they have learned from very few examples?

Another example is provided by the field of *machine learning*, the subfield of computer science which is about designing algorithms that can learn from what they observe. The main issue there is that machine learning models are typically trained on ridiculously small amount of data, compared with the wealth of information truly contained in the real world. Action recognition tools, for instance, are trained (and tested) on benchmark datasets that contain, at best, a few tens of thousands of videos – compare that with the billions of videos one can access on YouTube. How can we make sure that they learn the right lesson? Should they not aim to work with sets of models rather than precise models?

As we will see in Chapter 17, random set theory can provide more robust foundations for machine learning 'in the wild'.[10] Statistical learning theory [1849, 1851, 1850] derives generalisation bounds on the error committed by trained models on new test data by assuming that the training and test distributions are the same. In opposition, assuming that both distributions, while distinct, belong to a given random set allows us to compute bounds which are more robust to real-world situations – this concept is illustrated in Fig. 1.6.

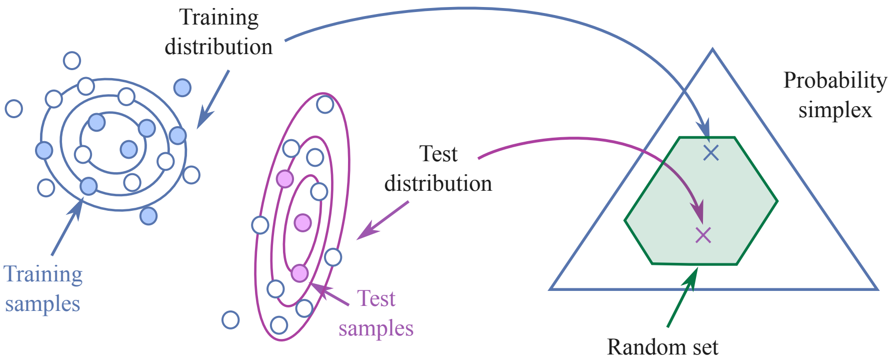

Fig. 1.6: A random-set generalisation of statistical learning theory, as proposed in Chapter 17.

Constraints on 'true' distributions From a statistical point of view, one can object that, even assuming that the natural description of the variability of phenomena is a probability distribution, under the law of large numbers probability distributions

[10] http://lcfi.ac.uk/news-events/events/reliable-machine-learning-wild/.

are the outcome of an infinite process of evidence accumulation, drawn from an infinite series of samples. In all practical cases, then, the available evidence may only provide some sort of constraint on the unknown, 'true' probability governing the process [1589]. Klir [988], among others, has indeed argued that 'imprecision of probabilities is needed to reflect the amount of information on which they are based. [This] imprecision should decrease with the amount of [available] statistical information.'

Unfortunately, those who believe probabilities to be limits of relative frequencies (the frequentists) never really 'estimate' a probability from the data – they merely assume ('design') probability distributions for their p-values, and test their hypotheses on them. In opposition, those who do estimate probability distributions from the data (the Bayesians) do not think of probabilities as infinite accumulations of evidence but as degrees of belief, and content themselves with being able to model the likelihood function of the data.

Both frequentists and Bayesians, though, seem to be happy with solving their problems 'asymptotically', thanks to the limit properties of maximum likelihood estimation, and the Bernstein–von Mises theorem's guarantees on the limit behaviour of posterior distributions. This hardly fits with current artificial intelligence applications, for instance, in which machines need to make decisions on the spot to the best of their abilities.

Logistic regression In fact, frequentists do estimate probabilities from scarce data when performing stochastic regression.

Logistic regression, in particular, allows us, given a sample $Y = \{Y_1, \ldots, Y_n\}$, $X = \{x_1, \ldots, x_n\}$, where $Y_i \in \{0, 1\}$ is a binary outcome at time i and x_i is the corresponding measurement, to learn the parameters of a conditional probability relation between the two, of the form

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \tag{1.4}$$

where β_0 and β_1 are two scalar parameters. Given a new observation x, (1.4) delivers the probability of a positive outcome $Y = 1$. Logistic regression generalises deterministic linear regression, as it is a function of the linear combination $\beta_0 + \beta_1 x$. The n trials are assumed independent but not equally distributed, for $\pi_i = P(Y_i = 1|x_i)$ varies with the index i (i.e., the time instant of collection): see Fig. 1.7.

The parameters β_0, β_1 of the logistic function (1.4) are estimated by the maximum likelihood of the sample, where the likelihood is given by

$$L(\beta|Y) = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{Y_i}.$$

Unfortunately, logistic regression appears inadequate when the number of samples is insufficient or when there are too few positive outcomes (1s) [970]. Also, inference by logistic regression tends to underestimate the probability of a positive outcome (see Section 1.3.8).

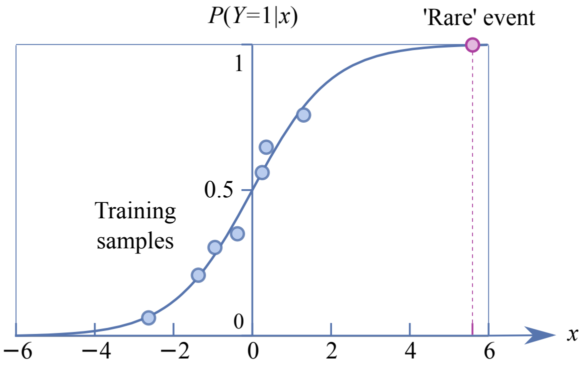

Fig. 1.7: Logistic regression and the notion of a ‘rare’ event.

Confidence intervals A major tool by which frequentists deal with the size of the sample is *confidence intervals*.

Let X be a sample from a probability $P(.|\theta,\phi)$ where θ is the parameter to be estimated and ϕ a nuisance parameter. A confidence interval for the parameter θ, with confidence level γ, is an interval $[u(X), v(X)]$ determined by the pair of random variables $u(X)$ and $v(X)$, with the property

$$\mathbb{P}(u(X) < \theta < v(X)|\theta,\phi) = \gamma \quad \forall(\theta,\phi). \tag{1.5}$$

For instance, suppose we observe the weight of 25 cups of tea, and we assume it is normally distributed with mean μ. Since the (normalised) sample mean Z is also normally distributed, we can ask, for instance, what values of the mean are such that $P(-z \leq Z \leq z) = 0.95$. Since $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$, this yields a (confidence) interval for μ, namely

$$P(\overline{X} - 0.98 \leq \mu \leq \overline{X} + 0.98).$$

Confidence intervals are a form of interval estimate. Their correct interpretation is about ‘sampling samples’: if we keep extracting new sample sets, 95% (say) of the time the confidence interval (which will differ for every new sample set) will cover the true value of the parameter. Alternatively, there is a 95% probability that the calculated confidence interval from some future experiment will encompass the true value of the parameter. One cannot claim, instead, that a specific confidence interval is such that it contains the value of the parameter with 95% probability.

A Bayesian version of confidence intervals also exists, under the name of *credible*[11] intervals.

1.3.8 Unusual data: Rare events

While the term ‘scarce data’ denotes situations in which data are of insufficient *quantity*, *rare events* [587] denotes cases in which the training data are of insufficient

[11] https://www.coursera.org/learn/bayesian/lecture/hWn0t/credible-intervals.

quality, in the sense that they do not properly reflect the underlying distribution. An equivalent term, coined by Nassim Nicholas Taleb, is ‘black swans’. This refers to an unpredictable event (also called a ‘tail risk’) which, once it has occurred, is (wrongly) rationalised in hindsight as being predictable/describable by the existing risk models. Basically, Knightian uncertainty is presumed not to exist, typically with extremely serious consequences. Examples include financial crises and plagues, but also unexpected scientific or societal developments. In the most extreme cases, these events may have never even occurred: this is the case for the question ‘will your vote will be decisive in the next presidential election?’ posed by Gelman and King in [673].

What does consitute a ‘rare’ event? We can say that an event is ‘rare’ when it covers a region of the hypothesis space which is seldom sampled. Although such events hardly ever take place when a single system is considered, they become a tangible possibility when very many systems are assembled together (as is the case in the real world). Given the rarity of samples of extreme behaviours (tsunamis, power station meltdowns, etc.), scientists are forced to infer probability distributions for the behaviour of these systems using information captured in ‘normal’ times (e.g. while a nuclear power plant is working just fine). Using these distributions to extrapolate results at the ‘tail’ of the curve via popular statistical procedures (e.g. logistic regression, Section 1.3.7) may then lead to sharply underestimating the probability of rare events [970] (see Fig. 1.7 again for an illustration). In response, Harvard’s G. King [970] proposed corrections to logistic regression based on oversampling rare events (represented by 1s) with respect to normal events (0s). Other people choose to drop generative probabilistic models entirely, in favour of discriminative ones [74]. Once again, the root cause of the problem is that uncertainty affects our very models of uncertainty.

Possibly the most straightforward way of explictly modelling second-order (Knightian) uncertainties is to consider sets of probability distributions, objects which go under the name of *credal sets*. In Chapter 17 we will show instead how belief functions allow us to model this model-level uncertainty, specifically in the case of logistic regression.

1.3.9 Uncertain data

When discussing how different mathematical frameworks cope with scarce or unusual data, we have implicitly assumed, so far, that information comes in the form of certainty: for example, I measure a vector quantity x, so that my conditioning event is $A = \{x\}$ and I can apply Bayes’ rule to update my belief about the state of the world. Indeed this is the way Bayes’ rule is used by Bayesians to reason (in time) when new evidence becomes available. Frequentists, on the other hand, use it to condition a parametric distribution on the gathered (certain) measurements and generate p-values (recall Section 1.2.3).

In many situations this is quite reasonable: in science and engineering, measurements, which are assumed to be accurate, flow in as a form of ‘certain’ (or so it is often assumed) evidence. Thus, one can apply Bayes' rule to condition a parametric model given a sample of such measurements $x_1, \ldots, x_T$ to construct likelihood functions (or p-values, if you are a frequentist).

Qualitative data In many real-world problems, though, the information provided cannot be put in a similar form. For instance, concepts themselves may be not well defined, for example 'this object is dark' or 'it is somewhat round': in the literature, this is referred to as *qualitative* data. Qualitative data are common in decision making, in which expert surveys act as sources of evidence, but can hardly be put into the form of measurements equal to sharp values.

As we will see in Section 6.5, *fuzzy theory* [2083, 973, 531] is able to account for not-well-defined concepts via the notion of *graded membership* of a set (e.g. by assigning every element of the sample space a certain degree of membership in any given set).

Unreliable data Thinking of measurements produced by sensor equipment as 'certain' pieces of information is also an idealisation. Sensors are not perfect but come with a certain degree of reliability. Unreliable sensors can then generate faulty (outlier) measurements: can we still treat these data as 'certain'? They should rather be assimilated to false statements issued with apparent confidence.

It then seems to be more sensible to attach a degree of reliability to any measurements, based on the past track record of the data-generating process producing them. The question is: can we still update our knowledge state using partly reliable data in the same way as we do with certain propositions, i.e., by conditioning probabilities via Bayes' rule?

Likelihood data Last but not least, evidence is often provided directly in the form of whole probability distributions. For instance, 'experts' (e.g. medical doctors) tend to express themselves directly in terms of chances of an event happening (e.g. 'diagnosis A is most likely given the symptoms, otherwise it is either A or B', or 'there is an 80% chance this is a bacterial infection'). If the doctors were frequentists, provided with the same data, they would probably apply logistic regression and come up with the same prediction about the conditional probability $P(\text{disease}|\text{symptoms})$: unfortunately, doctors are not statisticians.

In addition, sensors may also provide as output a probability density function (PDF) on the same sample space: think of two separate Kalman filters, one based on colour, the other on motion (optical flow), providing a Gaussian predictive PDF on the location of a target in an image.

Jeffrey's rule of conditioning Jeffrey's rule of conditioning [1690, 1227] is a step forward from certainty and Bayes' rule towards being able to cope with uncertain data, in particular when the latter comes in the form of another probability distribution. According to this rule, an initial probability P 'stands corrected' by a second probability P', defined only on a certain number of events.

Namely, suppose that P is defined on a σ-algebra $\mathcal{A}$, and that there is a new probability measure P' on a subalgebra $\mathcal{B}$ of $\mathcal{A}$.
If we require that the updated probability P''

1. has the probability values specified by P' for events in $\mathcal{B}$, and
2. is such that $\forall B \in \mathcal{B}$, $X, Y \subset B$, $X, Y \in \mathcal{A}$,

$$\frac{P''(X)}{P''(Y)} = \begin{cases} \frac{P(X)}{P(Y)} & \text{if } P(Y) > 0, \\ 0 & \text{if } P(Y) = 0, \end{cases}$$

then the problem has a unique solution, given by

$$P''(A) = \sum_{B \in \mathcal{B}} P(A|B) P'(B). \tag{1.6}$$

Equation (1.6) is sometimes also called the *law of total probability*, and obviously generalises Bayesian conditioning (obtained when $P'(B) = 1$ for some B).

Beyond Jeffrey's rule What if the new probability P' is defined on the same σ-algebra $\mathcal{A}$? Jeffrey's rule cannot be applied. As we have pointed out, however, this does happen when multiple sensors provide predictive PDFs on the same sample space.

Belief functions deal with uncertain evidence by moving away from the concept of *conditioning* (e.g., via Bayes' rule) to that of *combining* pieces of evidence simultaneously supporting multiple propositions to various degrees. While conditioning is an inherently asymmetric operation, in which the current state of the world and the new evidence are represented by a probability distribution and a single event, respectively, combination in belief function reasoning is completely symmetric, as both the current beliefs about the state of the world and the new evidence are represented by a belief function.

Belief functions are naturally capable of encoding uncertain evidence of the kinds discussed above (vague concepts, unreliable data, likelihoods), as well as of representing traditional 'certain' events. Qualitative concepts, for instance, are represented in the formalism by *consonant* belief functions (Section 2.8), in which the supported events are nested – unreliable measurements can be naturally portrayed as 'discounted' probabilities (see Section 4.3.6).

1.3.10 Knightian uncertainty

Second-order uncertainty is real, as demonstrated by its effect on human behaviour, especially when it comes to decision making. A classical example of how Knightian uncertainty empirically affects human decision making is provided by *Ellsberg's paradox* [565].

Ellsberg's paradox A *decision* problem can be formalised by defining:

- a set Ω of states of the world;
- a set $\mathcal{X}$ of consequences;
- a set $\mathcal{F}$ of acts, where an act is a function $f : \Omega \to \mathcal{X}$.

Let $\succcurlyeq$ be a *preference relation* on $\mathcal{F}$, such that $f \succcurlyeq g$ means that f is at least as desirable as g. Given $f, h \in \mathcal{F}$ and $E \subseteq \Omega$, let fEh denote the act defined by

$$(fEh)(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in E, \\ h(\omega) & \text{if } \omega \notin E. \end{cases} \tag{1.7}$$

Savage's *sure-thing principle* [1411] states that $\forall E, \forall f, g, h, h'$,

$$fEh \succcurlyeq gEh \Rightarrow fEh' \succcurlyeq gEh'.$$

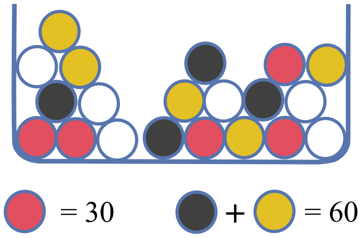

Fig. 1.8: Ellsberg's paradox.

Now, suppose you have an urn containing 30 red balls and 60 balls which are either black or yellow (see Fig. 1.8). Then, consider the following gambles:

- f_1: you receive 100 euros if you draw a red (R) ball;
- f_2: you receive 100 euros if you draw a black (B) ball;
- f_3: you receive 100 euros if you draw a red or a yellow (Y) ball;
- f_4: you receive 100 euros if you draw a black or a yellow ball.

In this example, $\Omega = \{R, B, Y\}$, $f_i : \Omega \to \mathbb{R}$ and $\mathcal{X} = \mathbb{R}$ (consequences are measured in terms of monetary returns). The four acts correspond to the mappings in the following table:

	R	B	Y
f_1	100	0	0
f_2	0	100	0
f_3	100	0	100
f_4	0	100	100

Empirically, it is observed that most people strictly prefer f_1 to f_2, while strictly preferring f_4 to f_3. Now, pick $E = \{R, B\}$. By the definition (1.7),

$$f_1\{R,B\}0 = f_1, \quad f_2\{R,B\}0 = f_2, \quad f_1\{R,B\}100 = f_3, \quad f_2\{R,B\}100 = f_4.$$

Since $f_1 \succcurlyeq f_2$, i.e., $f_1\{R,B\}0 \succcurlyeq f_2\{R,B\}0$, the sure-thing principle would imply that $f_1\{R,B\}100 \succcurlyeq f_2\{R,B\}100$, i.e., $f_3 \succcurlyeq f_4$. This empirical violation of the sure-thing principle is what constitutes the so-called Ellsberg paradox.

Aversion to 'uncertainty' The argument above has been widely studied in economics and decision making,[12] and has to do with people's instinctive aversion to (second-order) uncertainty. They favour f_1 over f_2 because the former ensures a guaranteed $\frac{1}{3}$ chance of winning, while the latter is associated with a (balanced) interval of chances between 0 and $\frac{2}{3}$. Although the average probability of success is still $\frac{1}{3}$, the lower bound is 0 – people tend to find that unacceptable.

Investors, for instance, are known to favour 'certainty' over 'uncertainty'. This was apparent, for instance, from their reaction to the UK referendum on leaving the European Union:

> "*In New York, a recent meeting of S&P Investment Advisory Services five-strong investment committee decided to ignore the portfolio changes that its computer-driven investment models were advising. Instead, members decided not to make any big changes ahead of the vote*."[13]

Does certainty, in this context, mean a certain outcome of an investor's gamble? Certainly not. It means that investors are confident that their models can fit the observed patterns of variation. In the presence of Knightian uncertainty, human beings assume a more cautious, conservative behaviour.

Climate change An emblematic application in which second-order uncertainty is paramount is climate change modelling. Admittedly, this constitutes an extremely challenging decision-making problem, where policy makers need to decide whether to invest billions of dollars/euros/pounds in expensive engineering projects to mitigate the effects of climate change, knowing that the outcomes of their decision will be apparent only in twenty to thirty years' time.

Rather surprisingly, the mainstream in climate change modelling is *not* about explicitly modelling uncertainty at all: the onus is really on developing ever more complex dynamical models of the environment and validating their predictions. This is all the more surprising as it is well known that even deterministic (but nonlinear) models tend to display chaotic behaviour, which induces uncertainty in predictions of their future state whenever initial conditions are not known with certainty. Climate change, in particular, requires making predictions very far off in the future: as dynamical models are obviously much simplified versions of the world, they become more and more inaccurate as time passes.

What are the challenges of modelling statistical uncertainty explicitly, in this context? First of all, the lack of priors (ouch, Bayesians!) for the climate space,

[12] `http://www.econ.ucla.edu/workingpapers/wp362.pdf`.

[13] `http://www.wsj.com/articles/global-investors-wake-up-to-brexit-threat\ -1466080015`.

whose points are very long vectors whose components are linked by complex dependencies. Data are also relatively scarce, especially as we go back in time: as we just saw, scarcity is a source of Knightian uncertainty as it puts constraints on our ability to estimate probability distributions. Finally, hypothesis testing cannot really be used either (too bad, frequentists!): this is clearly not a designed experiment where one can make sensible assumptions about the underlying data-generating mechanism.

1.4 Mathematics (plural) of uncertainty

It is fair to summarise the situation by concluding that something is wrong with both Kolmogorov's mathematical probability and its most common interpretations.

As discussed in our Preface, this realisation has led many authors to recognise the need for a mathematical theory of uncertainty capable of overcoming the limitations of classical mathematical probability. While the most significant approaches have been briefly recalled there, Chapter 6 is devoted to a more in-depth overview of the various strands of uncertainty theory.

1.4.1 Debate on uncertainty theory

Authors from a variety of disciplines, including, statistics, philosophy, cognitive science, business and computer science, have fuelled a lively debate [890] on the nature and relationships of the various approaches to uncertainty quantification, a debate which to a large extent is still open.

Back in 1982, a seminal paper by psychologists Kahneman and Tversky [943] proposed, in stark contrast to formal theories of judgement and decision, a formalisation of uncertainty which contemplates different variants associated with frequencies, propensities, the strength of arguments or direct experiences of confidence.

Bayesian probability: Detractors and supporters A number of authors have provided arguments for and against (Bayesian) probability as the method of choice for representing uncertainty.

Cheeseman, for instance, has argued that probability theory, when used correctly, is sufficient for the task of reasoning under uncertainty [255] ('In defense of probability'), while advocating the interpretation of probability as a measure of belief rather than a frequency ratio. In opposition, in his 'The limitation of Bayesianism' [1893], Wang has pointed out a conceptual and notational confusion between the explicit and the implicit condition of a probability evaluation, which leads to seriously underestimating the limitations of Bayesianism. The same author [1892] had previously argued that psychological evidence shows that probability theory is not a proper descriptive model of intuitive human judgement, and has limitations even as a normative model. A new normative model of judgement under uncertainty was then designed under the assumption that the system's knowledge and resources are insufficient with respect to the questions that the system needs to answer.

In his response to Shafer's brief note [1590] on Lindley's paradox (see Section 1.2.6), Lindley himself [426] argued that the Bayesian approach comes out of Shafer's criticism quite well. Indeed, Lindley's paradox has been the subject of a long list of analyses [828, 706, 1828].

More recently (2006), Forster [635] has considered from a philosophical perspective what he called the likelihood theory of evidence [87], which claims that all the information relevant to the bearing of data on hypotheses is contained in the likelihoods, showing that there exist counter-examples in which one can tell which of two hypotheses is true from the full data, but not from the likelihoods alone. These examples suggest that some forms of scientific reasoning, such as the consilience of inductions [1619], cannot be represented within the Bayesian and likelihoodist philosophies of science.

Relations between uncertainty calculi A number of papers have presented attempts to understand the relationships between apparently different uncertainty representations. In an interesting 1986 essay, for instance, Horvitz [841] explored the logical relationship between a small number of intuitive properties for belief measures and the axioms of probability theory, and discussed its relevance to research on reasoning under uncertainty.

In [822], Henkind analysed four of (what were then) the more prominent uncertainty calculi: Dempster–Shafer theory, fuzzy set theory, and the early expert systems MYCIN [200, 778] and EMYCIN. His conclusion was that there does not seem to be one calculus that is the best for all situations. Other investigators, including Zimmerman, have supported an application-oriented view of modelling uncertainty, according to which the choice of the appropriate modelling method is context dependent. In [2132], he suggested an approach to selecting a suitable method to model uncertainty as a function of the context. Pearl's 1988 survey of evidential reasoning under uncertainty [1401] highlighted a number of selected issues and trends, contrasting what he called *extensional* approaches (based on rule-based systems, in the tradition of classical logic) with *intensional* frameworks (which focus on 'states of the world'), and focusing on the computational aspects of the latter methods and of belief networks of both the Bayesian and the Dempster–Shafer type.

Dubois and Prade [540] pointed out some difficulties faced by non-classical probabilistic methods, due to their relative lack of maturity. A comparison between the mathematical models of expert opinion pooling offered by Bayesian probabilities, belief functions and possibility theory was carried out, proving that the Bayesian approach suffers from the same numerical stability problems as possibilistic and evidential rules of combination in the presence of strongly conflicting information. It was also suggested that possibility and evidence theories may offer a more flexible framework for representing and combining subjective uncertain judgements than the framework of subjective probability alone.

Kyburg [1088] also explored the relations between different uncertainty formalisms, advocating that they should all be thought of as special cases of sets of probability functions defined on an algebra of statements. Thus, interval probabilities should be construed as maximum and minimum probabilities within a set of distributions, belief functions should be construed as lower probabilities, etc.

Philippe Smets [1702], on his side, surveyed various forms of imperfect data, classified into either imprecision, inconsistency or uncertainty. He argued that the greatest danger in approximate reasoning is the use of inappropriate, unjustified models, and made a case against adopting a single model and using it in all contexts (or, worse, using all models in a somewhat random way) [1684]. The reason is that ignorance, uncertainty and vagueness are really different notions which require different approaches. He advised that, before using a quantified model, we should:

1. Provide canonical examples justifying the origin of the numbers used.
2. Justify the fundamental axioms of the model and their consequences, via 'natural' requirements.
3. Study the consequence of the derived models in practical contexts to check their validity and appropriateness.

A common error, he insisted, consists in accepting a model because it 'worked' nicely in the past, as empirical results can only falsify a model, not prove that it is correct.

Approximate reasoning Several papers have dealt with the issue of uncertainty in the context of artificial intelligence, expert systems or, as it is sometimes referred to, *approximate reasoning*.

In a 1987 work, Shafer [1597] discussed the challenges arising from the interaction of artificial intelligence and probability, identifying in particular the issue of building systems that can design probability arguments. Thompson [1816], after acknowledging that there was no general consensus on how best to attack evidential reasoning, proposed a general paradigm robust enough to be of practical use, and used it to formulate classical Bayes, convex Bayes, Dempster–Shafer, Kyburg and possibility approaches in a parallel fashion in order to identify key assumptions, similarities and differences. Ruspini [1512] argued in 1991 that approximate reasoning methods are sound techniques that describe the properties of a set of conceivable states of a real-world system, using a common framework based on the logical notion of 'possible worlds'. In his 1993 book, Krause [1067] supported the view that an eclectic approach is required to represent and reason under the many facets of uncertainty. Rather than the technical aspects, that book focuses on the foundations and intuitions behind the various schools. Chapter 4 of it, 'Epistemic probability: The Dempster–Shafer theory of evidence', is devoted entirely to belief theory.

The recent debate A more recent publication by Castelfranchi et al. [235], has focused on the central role of 'expectations' in mental life and in purposive action, reducing them in terms of more elementary ingredients, such as beliefs and goals. There, the authors allow the possibility that beliefs in a proposition and its negation do not add up to one, as in the belief function framework.

Gelman [672] has pointed out the difficulties with both robust Bayes and belief function approaches, using a simple example involving a coin flip and a boxing/wrestling match. His conclusions are that robust Bayes approaches allow ignorance to spread too broadly, whereas belief functions inappropriately collapse to simple Bayesian models.

Keppens [956] has argued that the use of subjective probabilities in evidential reasoning (in particular for crime investigation) is inevitable for several reasons, including lack of data, non-specificity of phenomena and fuzziness of concepts in this domain. His paper argues that different approaches to subjective probability are really concerned with different aspects of vagueness.

1.4.2 Belief, evidence and probability

As recalled in the Preface, this book mostly focuses on the theory of belief functions, one of the most widely adopted formalisms for a mathematics of uncertainty, its geometric interpretation and its links with other uncertainty theories.

Belief functions as random sets The notion of a belief function originally derives from a series of seminal publications [415, 417, 418] by Arthur Dempster on upper and lower probabilities induced by multivalued mappings. Given a probability distribution p on a certain sample space, and a one-to-many map from such a sample space to another domain, p induces a probability distribution (a *mass assignment*) on the power set of the latter [415], i.e., a *random set* [1268, 1344]. A very simple example of such a mapping was given in the cloaked die example (Figure 1.4).

The term *belief function* was coined by Glenn Shafer [1583], who proposed to adopt these mathematical objects to represent evidence in the framework of subjective probability, and gave an axiomatic definition of them as non-additive (indeed, superadditive) probability measures. As mentioned when recalling Jeffrey's rule (Section 1.3.9), in belief theory conditioning (with respect to an event) is replaced by combination (of pieces of evidence, represented by belief functions).

As a result, as we will see in more detail in Part I, the theory of belief functions addresses all the issues with the handling of uncertainty we discussed in this Introduction. It does not assume an infinite amount of evidence to model imprecision, but uses all the available partial evidence, coping with missing data in the most natural of ways. It properly represents ignorance by assigning mass to the whole sample space or 'frame of discernment', and can coherently represent evidence on different but compatible domains. Furthermore, as a straightforward generalisation of probability theory, its rationale is rather neat and does not require us to entirely abandon the notion of an event (as opposed to Walley's imprecise probability theory [1874]), although it can be extended to assign basic probabilities to real-valued functions [153, 1078] rather than events. Finally, it contains as special cases both fuzzy set theory and possibility theory.

Belief theory as evidential probability Shafer called his 1976 proposal [1583] 'A mathematical theory of evidence', whereas the mathematical objects it deals with are termed 'belief functions'. Where do these names come from, and what interpretation of probability (in its wider sense) do they entail?

In fact, belief theory is a theory of epistemic probability: it is about probabilities as a mathematical representation of knowledge (never mind whether it is a human's knowledge or a machine's). *Belief* is often defined as a state of mind in which a person thinks something to be the case, with or without there being empirical evidence in support. *Knowledge* is a rather more controversial notion, for it is regarded by some as the part of belief that is true, while others consider it as that part of belief which is *justified* to be true. *Epistemology* is the branch of philosophy concerned with the theory of knowledge. *Epistemic probability* (Fig. 1.9) is the study of probability as a representation of knowledge.

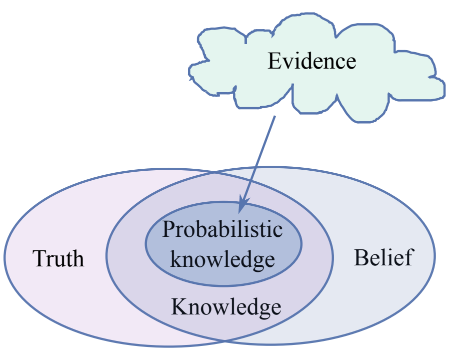

Fig. 1.9: Belief function theory is an instance of evidential, epistemic probability.

The theory of evidence is also, as the name itself suggests, a theory of *evidential* probability: one in which the probabilities representing one's knowledge are induced ('elicited') by the evidence at hand. In probabilistic logic [586, 205], statements such as 'hypothesis H is probably true' are interpreted to mean that the empirical evidence E supports hypothesis H to a high degree – this degree of support is called the *epistemic probability* of H given E. As a matter of fact, Pearl and others [1405, 1682, 137] have supported a view of belief functions as probabilities on the logical causes of a certain proposition (the so-called *probability of provability* interpretation), closely related to modal logic [800].

The rationale for belief function theory can thus be summarised as follows: there exists evidence in the form of probabilities, which supports degrees of belief on the matter at hand. The space where the (probabilistic) evidence lives is different from the hypothesis space (where belief measures are defined). The two spaces are linked by a one-to-many map, yielding a mathematical object known as a random set [415]. In Chapter 2 we will recall the basic elements of the theory of evidence and the related mathematical definitions.