Data-Based Decisions under Complex Uncertainty

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik der
Ludwig-Maximilians-Universität München

Robert Hable

03. November 2008

Erstgutachter:	Prof. Dr. Thomas Augustin
Zweitgutachter:	PD Dr. Christian Heumann
Drittgutachter:	RNDr. Jiřina Vejnarová, CSc.
Rigorosum:	05. Februar 2009

Abstract

Decision theory is, in particular in economics, medical expert systems and statistics, an important tool for determining optimal decisions under uncertainty. In view of applications in statistics, the present book is concerned with decision problems which are explicitly data-based. Since the arising uncertainties are often too complex to be described by classical precise probability assessments, concepts of imprecise probabilities (coherent lower previsions, F-probabilities) are applied. Due to the present state of research, some basic groundwork has to be done: Firstly, topological properties of different concepts of imprecise probabilities are investigated. In particular, the concept of coherent lower previsions appears to have advantageous properties for applications in decision theory. Secondly, several decision theoretic tools are developed for imprecise probabilities. These tools are mainly based on concepts developed by L. Le Cam and enable, for example, a definition of sufficiency in case of imprecise probabilities for the first time.

Building on that, the article [A. Buja, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65 (1984) 367-384] is reinvestigated in the only recently available framework of imprecise probabilities. This leads to a generalization of results within the Huber-Strassen theory concerning least favorable pairs or models.

Results obtained by these investigations can also be applied afterwards in order to justify the use of the method of natural extension, which is fundamental within the theory of imprecise probabilities, in data-based decision problems. It is shown by means of the theory of vector lattices that applying the method of natural extension in decision problems does not affect the optimality of decisions. However, it is also shown that, in general, the method of natural extension suffers from a severe instability.

The book closes with an application in statistics in which a minimum distance estimator is developed for imprecise probabilities. After an investigation concerning its asymptotic properties, an algorithm for calculating the estimator is given which is based on linear programming. This algorithm has led to an implementation of the estimator in the programming language R which is publicly available as R package “imprProbEst”. The applicability of the estimator (even for large sample sizes) is demonstrated in a simulation study.

Zusammenfassung

Die Entscheidungstheorie ist vor allem in den Wirtschaftswissenschaften, bei medizinischen Expertensystemen und in der Statistik ein wichtiges Werkzeug zur Bestimmung optimaler Entscheidungen in Unsicherheitssituationen. Im Hinblick auf statistische Anwendungen beschäftigt sich die vorliegende Arbeit mit Entscheidungsproblemen, die explizit datenbasiert sind. Da die auftretenden Unsicherheiten häufig zu komplex sind, um durch klassische präzise Wahrscheinlichkeiten beschrieben werden zu können, werden Konzepte unscharfer Wahrscheinlichkeiten (coherent lower previsions, F-Wahrscheinlichkeiten) verwendet. Entsprechend des derzeitigen Forschungsstands sind zunächst einige Grundlagenarbeiten nötig: Zum einen werden topologische Eigenschaften verschiedener Konzepte unscharfer Wahrscheinlichkeiten untersucht, wobei sich herausstellt, dass vor allem coherent lower previsions günstige Eigenschaften für die Verwendung in der Entscheidungstheorie besitzen. Zum anderen werden verschiedene entscheidungstheoretische Werkzeuge speziell für die Situation unscharfer Wahrscheinlichkeiten entwickelt. Diese basieren hauptsächlich auf Arbeiten von L. Le Cam und ermöglichen zum Beispiel erstmals eine Definition der Suffizienz für unscharfe Wahrscheinlichkeiten.

Aufbauend auf diese Grundlagen wird die Arbeit [A. Buja, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65 (1984) 367-384] mit Hilfe der noch relativ neuen Konzepte unscharfer Wahrscheinlichkeiten untersucht. Hierdurch ergibt sich eine Verallgemeinerung von Resultaten aus der Huber-Strassen-Theorie über ungünstigste Paare bzw. Modelle.

Dabei gewonnene Resultate werden anschließend dazu verwendet, um die für die Theorie der coherent lower previsions grundlegende Methode der natural extension für die Verwendung bei datenbasierten Entscheidungsproblemen zu rechtfertigen. Es wird mit Hilfe der Vektorverbandstheorie gezeigt, dass Anwendungen der natural extension in Entscheidungsproblemen keinen Einfluss auf die Optimalität von Entscheidungen haben. Andererseits wird aber auch gezeigt, dass – ganz unabhängig von der Entscheidungstheorie – die Methode der natural extension ein ernstes Stabilitätsproblem besitzt.

Als statistische Anwendung wird abschließend ein Minimum-Distanz-Schätzer für unscharfe Wahrscheinlichkeiten entwickelt und auf dessen asymptotische Eigenschaften untersucht. Basierend auf linearer Programmierung wird ein Algorithmus zur Berechnung des Schätzers erarbeitet, der inzwischen in der Programmiersprache R implementiert und als R-Paket “imprProbEst” frei verfügbar ist. Die Anwendbarkeit des Schätzers (auch bei großen Stichprobenumfängen) wird in einer Simulationsstudie demonstriert.

Acknowledgments

First of all, I would like to thank my advisor, Thomas Augustin, for his guidance, suggestions and support – not only with respect to this work. I am particularly grateful to him for giving me a great amount of freedom in my research. Writing a Ph.D. thesis under his guidance is a pleasure. He also had a very positive influence on my style of writing as he convinced me (and frequently reminded me) to add some text between the formulas.

Next, I would like to thank Andreas Christmann and Helmut Rieder who gave me the opportunity to finish my Ph.D. thesis in Bayreuth. In addition, I thank Helmut Rieder and Peter Ruckdeschel for drawing my attention to the work of A. Buja and L. Le Cam which had a high impact on the present book. I am also much indepted to Matthias Kohl for his support in programming with R which enabled me to add a simulation study.

Last but not least, I would like to thank the Cusanuswerk (foundation of the Roman Catholic Church) for a Ph.D. scholarship; the present work would not have been possible without it.

Contents

Chapter 1

Introduction

Every day life permanently forces to make decisions. Though most of these decisions are made without deep and sophisticated considerations (and especially not based on advanced mathematical evaluations), there are many situations where such simple decision making is not possible or advisable. If the consequences of decisions are certain, decision making is no problem even in complex situations. However, consequences usually happen to be uncertain and, therefore, the wish to be able to make good decisions in such situations gave rise to the development of a theory about "how to make good decisions" – called decision theory. Such a theory is essential today the more so as modern civilization leads to increasing complexity in many fields where single decision makers cannot have enough overview over all possibilities, chances and risks in order to make good and traceable decisions. Therefore, decision theory has become a very popular tool for determining good decisions under uncertainty which has been successfully applied especially in economics and medical expert systems.

A current example of a challenging issue where decisions have to be made is climate change. Usually, uncertainties are modeled by probabilities in decision theory but, here, the uncertainties are much too complex so that precise probabilities cannot be given. This is because the present state of research is far away from providing sufficient insight into all relevant facts. The high degree of uncertainty with respect to climate change is for example demonstrated in Held et al. (2008) which reports on a survey among some experts. Nevertheless, decisions have to be made now. In such complex situations, uncertainties cannot be adequately modeled by ordinary precise probabilities since two different kinds of uncertainty are involved. The first one is the uncertainty about the outcome of a random event – this can suitably be modeled by usual probabilities. The second one is the uncertainty about the random process itself – that is, the uncertainty about the "true" probabilities. This latter kind of uncertainty is often called *ambiguity* and has already attracted considerable attention, in particular in artificial intelligence[1] and in economics after the publication of the seminal article Ellsberg (1961) (cf. e.g. Hamouda and Rowley (1997) which contains a large collection of articles influenced by Ellsberg (1961)). In this article, D. Ellsberg shows in an experiment carried out with economists that, in case of ambiguity, "reasonable" decision making can be in complete contradiction to classical decision theory where all uncertainties are modeled by precise probabilities. This so-called Ellsberg paradox demonstrates that, in particular in decision

[1] Cf. the proceedings of the annual "Conference on Uncertainty in Artificial Intelligence", e.g. McAllester and Myllymäki (2008), see also the Homepage of the "Association for Uncertainty in Artificial Intelligence (AUAI)" in the Internet: www.auai.org.

theory, ambiguity plays a constitutive role. Therefore, a formalization of ambiguity is needed, the more so as ambiguity seems to be rather the rule than the exception in real decision problems.
Such formalizations have been developed under the name *imprecise probabilities*. Since there are several different formalizations of ambiguity, the theory of imprecise probabilities is rather a collection of different theories than one single theory; the present book is concerned with two of the most important concepts of imprecise probabilities, namely coherent lower previsions (Walley, 1991) and interval probabilities (Weichselberger, 2001). Another important concept is e.g. the Dempster-Shafer theory of belief functions (e.g. Yager et al. (1994)). [2] The theory of imprecise probabilities states that probabilities can rarely be specified by precise numbers in real applications. For example, assuming the standard normal distribution means that the probability of observing a value greater than 1 would exactly be equal to

$$P\big((1,\infty)\big) \;=\; 0.1586553\ldots$$

Of course, such a precise assumption is unrealistic and it would be much more realistic to assume that, e.g. the above probability lies somewhere between 0.12 and 0.17:

$$\underline{P}\big((1,\infty)\big) \;=\; 0.12 \;\leq\; P\big((1,\infty)\big) \;\leq\; 0.17 \;=\; \overline{P}\big((1,\infty)\big)$$

That is, only lower and upper bounds are given instead of precise probabilities. Such a generalization of ordinary probability theory, in fact, offers a suitable formalization of complex uncertainties where ambiguity is involved: Uncertainty in form of random processes is modeled by the precise probabilities which are in accordance with the lower and upper bounds; uncertainty in form of ambiguity is modeled by the bandwidth of possible probabilities which are in accordance with the lower and upper bounds – the larger the bandwidth is, the greater ambiguity is.[3] Accordingly, Weichselberger and Augustin (1998) shows that the Ellsberg paradox can be solved if all uncertainties are modeled by imprecise probabilities.

Concepts of imprecise probabilities have already been used in many decision theoretic investigations. For example, a classical text in mathematical economics is Gilboa and Schmeidler (1989) and there are also a number of recent articles in mathematical finance where imprecise probabilities (and equivalent concepts) are applied in a decision theoretic setup; confer e.g. Schied (2006), Maccheroni et al. (2006) and Föllmer et al. (2007) where the latter article provides an overview including many references.[4] [5] In the above mentioned example concerning climate change, imprecise probabilities are used for decision theoretic evaluations e.g. by Kriegler (2005) and Hall et al. (2007). The topic of climate change has also been used in an experiment concerning psychological aspects of decision making with imprecise probabilities; see Budescu et al. (2008).

[2] The theory of fuzzy sets (e.g. Zadeh and Kacprzik (1992)) is also thematically related to the theory of imprecise probabilities.

[3] Complete ignorance, for example, can be modeled by the bounds

$$\underline{P}(A) \;=\; 0 \qquad \text{and} \qquad \overline{P}(A) \;=\; 1$$

[4] Ambiguity is often associated with the term "Knightian Uncertainty" in economics.

[5] The concept of imprecise probabilities due to Walley (1991) is also mathematically equivalent to the systematic approach to risk measures in mathematical finance provided by the influential article Artzner et al. (1999); cf. Vigic (2008).

Most parts of this book use the concept of imprecise probabilities due to Walley (1991) which is called “coherent lower prevision” [6]. A general article about decision making with coherent lower previsions is de Cooman and Walley (2002). Different optimality criteria are discussed by Schervish et al. (2003) and Troffaes (2007) in this setup – choosing a suitable optimality criterion is fundamental for the whole theory because this determines the meaning of the notions “good decision” and “optimal decision”. Algorithms for the calculation of optimal decisions are given by Kikuti et al. (2005) and Utkin and Augustin (2005).

A serious limitation of the present state of research is that most investigations in this setup are only concerned with *data-free* decision problems while, in most real decision problems, decision making can be based on data. For example, a company which has to decide about a certain investment can look at suitable economic indicators such as stock market prices or retail sales. In the classical setup where all kinds of uncertainties are tried to be modeled by precise probabilities, investigations of decision problems which are explicitly data-based are not necessary because the main theorem of Bayesian decision theory (cf. (Berger, 1985, § 4.4.1)) states that every data-based decision problem can be solved by solving a corresponding data-free decision problem. However, Augustin (2003) shows that an analog statement is not true in case of imprecise probabilities. Therefore, data-based decision problems have to be considered as a matter of its own when dealing with imprecise probabilities – at least from a frequentist point of view.[7]

One of the most important fields of application is statistics where decisions (e.g. rejecting a hypothesis in hypothesis testing) are always data-based. As described by Wald (1950), statistics can be formalized as a special case of decision theory and this discovery has led to an own area of research called statistical decision theory. Accordingly, many books about mathematical statistics are written in terms of decision theory, for example Berger (1985), Strasser (1985), Le Cam (1986), Liese and Miescke (2008).

In contrast to most decision theoretic evaluations under imprecise probabilities, the present book is concerned with decision problems (under imprecise probabilities) which are explicitly data-based; special interest lies in applications in mathematical statistics. While the above mentioned elaborated concepts of imprecise probabilities have only recently been developed, robust statistics already has successfully been dealing with a special case of ambiguity – namely small (but unknown) deviations from an ideal statistical model – in statistics for several decades. Usually, a statistician is faced with a set of hypotheses Θ and he is expected to draw conclusions about the unknown true hypothesis $\theta_0 \in \Theta$. In particular, this is possible if the statistician knows a family of probability measures

$$P_\theta\,, \qquad\qquad \theta \in \Theta$$

on a sample space $(\Xi, \mathcal{B})$ and if data $x_1, \ldots, x_n$ are available which are distributed according to the probability measure P_{θ_0}. Roughly speaking, the correctness of a hypothesis $\theta \in \Theta$ is the more plausible the better the observations are in line with P_θ . In spite of the often arising complexity of uncertainties, assuming that data precisely stem from one of the following few distributions is extremely popular in statistical evaluations: binomial, Poisson, uniform, normal, exponential, chi-square, t- and F-distribution. Though the use

[6] or equivalently “coherent upper prevision”

[7] According to Augustin (2003), the question if data-based decision problems have to be considered explicitly picks up an old debate between frequentists and Bayesians: Does the posterior distribution contain all relevant information after observing the data x?

of these distributions often seems to be based on tradition, there are also strong objective reasons for the use of these distributions in many situations. For example, the use of the normal distribution can often be justified by the central limit theorem. Nevertheless, it would be a misuse of the central limit theorem to assume that the data exactly stem from a normal distribution – this theorem only justifies the assumption that the data *approximately* stem from a normal distribution. Since models are usually not intended to exactly reflect the real world, one might argue that the normal distribution serves as a good approximation of reality and, therefore, leads to approximately correct results. However, it has already been known for many decades that this is simply not correct in general. Imperceptible small deviations from the ideal model (e.g. consisting of normal distributions) may lead to arbitrarily wrong conclusions; cf. e.g. Tukey (1960), Huber (1965), (Huber, 1981, § 1.1), (Hampel et al., 1986, § 1.2), (Marazzi, 1993, Introduction), (Huber, 1997, § 1) and (Kohl, 2005, Introduction).
Therefore, robust statistics aims to develop statistical procedures which are insensitive to (small) deviations from an ideal model.[8] To this end, it is often assumed that the data are approximately distributed according to an ideal model – the above mentioned parametric models commonly serve as such ideal models. Accordingly, (Hampel et al., 1986, p. 7) suggests the following definition of robust statistics:

> "Robust statistics as a collection of related theories, is the statistics of approximate parametric models."

Though assuming a known ideal parametric model can often be justified e.g. by use of the central limit, this is not always possible and, then, applying imprecise probabilities may help because these are intended to deal also with more general kinds of ambiguity.

An important fact about imprecise probabilities is that they can often be interpreted in various ways; cf. (Walley, 1991, § 2.10) and (Weichselberger, 2001, § 1.4 and § 1.5). Sometimes, the way of interpreting imprecise probabilities affects the definition of concepts which are generalized from classical probability theory to imprecise probabilities.[9] In this case, the sensitivity analyst's point of view is adopted so that the evaluations are in accordance with robust statistics. In short, this means that the observations are assumed to be distributed according to a precise probability distribution but this precise distribution is only imprecisely known.[10]

Since the concepts of imprecise probabilities due to Walley (1991) (coherent upper previsions) and Weichselberger (2001) (F-probabilities) as well are in accordance with this point of view, both concepts could be applied in the following treatment of decision theory under imprecise probabilities. Therefore, Chapter 2 takes a closer look on these concepts and compares them to each other. This comparison is especially concerned with those properties which are important in decision theoretic evaluations.
In order to do such a comparison, it is necessary to reformulate the definition and some basic properties of coherent upper previsions in terms of sample spaces, measurable functions and (finitely additive) probability measures since Walley (1991) does hardly use these terms but is written in terms of gambles, buying prices and previsions. This reformulation also makes it possible to apply, generalize and refer to concepts already used in

[8] Confer for example the following books about robust statistics: Huber (1981), Hampel et al. (1986), Rieder (1994), Jurečková and Sen (1996), Müller (1997), Wilcox (1997), Jurečková and Picek (2006) and Maronna et al. (2006).

[9] Cf. (Walley, 1991, § 2.10.5).

[10] Cf. (Walley, 1991, § 2.10.4).

classical probability theory or statistics. As a link between the two investigated concepts of imprecise probabilities, a third concept is presented which is called "upper expectations" and lies between Walley's and Weichselberger's concept. These upper expectations have originally been defined by Buja (1984) within robust statistics and have not been considered within the theory of imprecise probabilities so far.

Most part of Chapter 2 is concerned with topology. This is because, for the use of imprecise probabilities in decision theory, it is crucial that they have certain compactness properties. Compactness enables the application of minimax theorems which are most important in decision theory and, in particular, in decision theory under imprecise probabilities. In this way, the results of Section 2.3 and Section 2.4 yields that, with respect to decision theoretic evaluations, coherent upper previsions have more appropriate topological properties than upper expectations and F-probabilities. As a consequence of these investigations, the concept of coherent upper previsions is mostly used throughout this book.

Though coherent upper previsions and upper expectations are indeed different concepts (upper expectations may be seen as special cases of coherent upper previsions), they coincide in a very abstract sense: It is shown in Section 2.5 that every coherent upper prevision can be represented by an upper expectation via a Stone representation. Despite of the abstractness of this result, there is an important application of such representations which is successfully used later on in Subsection 3.3.3: In case of coherent upper previsions, σ-additivity is not necessarily available so that it is often hard or even impossible to generalize concepts from classical probability theory which rely on σ-additivity. Now, the above representation offers a canonical way for such definitions because σ-additivity is at hand for upper expectations.

While Chapter 2 is only concerned with imprecise probabilities, Chapter 3 turns over to data-based decision theory under coherent upper previsions. After some basic definitions and a description of the setup in Section 3.2, some more extended decision theoretic concepts are developed in Section 3.3. Most of these concepts are reformulations or extensions of concepts which have originally been defined by L. Le Cam in case of precise probabilities and have not been used in case of imprecise probabilities before. The definition of generalized randomizations in Subsection 3.3.1 is most fundamental. These generalized randomizations serve as a main tool in the present book, and the definitions of the other extended decision theoretic tools are based on this concept. Generalized randomizations [11] have been defined by L. Le Cam in order to be able to deal with large models which are not dominated by σ-finite measures [12]. Since imprecise probabilities consisting of coherent upper previsions are, in fact, very large models, the concept of generalized randomizations proves to be very appropriate when dealing with imprecise probabilities.

Another concept developed in Chapter 3 is "sufficiency", which is a very important concept in classical statistics and has not been defined in case of imprecise probabilities before. Since the most common definition of sufficiency is heavily based on conditional probabilities and the definition of *imprecise* conditional probabilities is still a matter of research (cf. e.g. de Cooman (2001), Weichselberger and Augustin (2003) and Škulj (2006)), defining sufficiency in case of imprecise probabilities might have been considered as being out of the scope of present research. However, there is also an alternate way of defining sufficiency which essentially goes back to Blackwell (1951) and has been generalized in Le Cam (1964). This definition is not formulated in terms of conditional probabilities

[11] also called transitions by L. Le Cam

[12] that is, if it is not possible to solely work with probability densities

but in terms of randomizations so that a generalization in case of imprecise probabilities is possible. The definition of sufficiency for imprecise models given in Subsection 3.3.2 is an extension of the notion "worst-case sufficiency" defined for upper expectations in Buja (1984).

As the concepts introduced in Section 3.3 are strongly connected with concepts introduced by L. Le Cam, Chapter 3 closes with Section 3.4 where the connections to Le Cam's setup are explained. On the one hand, Le Cam's setup is more specific than the setup used in imprecise probabilities because Le Cam only deals with precise probabilities. On the other hand, his setup is more general because he does not consider explicitly specified sample spaces but considers probabilities as elements of certain vector lattices. Furthermore, Section 3.4 may also serve as an introduction to Le Cam's abstract setup.

Since decision problems which are explicitly data-based have rarely been considered within the theory of imprecise probabilities so far, Chapters 2 and 3 have to develop some fundamentals and general tools which are intended to provide a base for subsequent theoretical studies. The remaining chapters of the present book turn over to some investigations in data-based decision theory under imprecise probabilities which seem to be most urgent: Firstly, the seminal article Buja (1984) has to be revised within the theory of imprecise probabilities in Chapter 4 because this early work is concerned with data-based decision theory under *upper expectations* and, therefore, is concerned with almost the same setup than the present book – even though Buja (1984) was published several years before Walley (1991) and Weichselberger (2001). Secondly, Chapter 5 is about the method of natural extension. Though this method is fundamental in the theory of imprecise probabilities due to Walley (1991), its use in data-based decision theory still needs some justification. Finally, Chapter 6 provides an application in statistical decision theory. There, a minimum distance estimator is developed and applied in a simulation study from which it can be seen that it is, in fact, possible to develop theoretically well-founded procedures in imprecise probabilities which can be used in applications. This meets objections that, due to high computational costs, imprecise probabilities could not be used for practical purposes. As a side-effect, Chapters 4, 5 and 6 demonstrate that the foundations and concepts developed in Chapters 2 and 3 are expedient for investigations in data-based decision theory under imprecise probabilities.

The remaining part of this introduction provides a more detailed overview of Chapters 4, 5 and 6:

Like Buja (1984), Chapter 4 is concerned with least favorable models – a topic which has received much attention in a special case of our setup namely robust hypothesis testing. Least favorable models are a matter of particular interest because direct solutions of decision problems under imprecise probabilities are quite often computationally intractable but, in the presence of least favorable models, computationally tractable solutions may be possible.

Most of the research concerning least favorable models was encourage by the celebrated article Huber and Strassen (1973). This article deals with hypothesis testing where a (rather special) upper prevision is tested against another one. There, it is shown that testing between these imprecise probabilities can be reduced to a testing problem between certain "least favorable" precise probabilities Q_0 and Q_1 . The pair (Q_0, Q_1) is called "least favorable pair" then. In this way, the computational effort of the original testing problem can often be reduced substantially. Least favorablility has attracted enormous attention

after the publication of Huber and Strassen (1973); see e.g. Rieder (1977), Österreicher (1978) and Hafner (1992); a detailed review of Huber and Strassen (1973) and the work following Huber and Strassen (1973) is given by Augustin (2002). In quite general data-based decision theory, where there are n states of nature (instead of two as in hypothesis testing), Buja (1984) develops the concept of least favorable models which generalizes the concept of least favorable pairs. There, a necessary and sufficient condition for the existence of such least favorable models is given. As already mentioned above, imprecise probabilities are modeled by upper expectations in Buja (1984).
The investigations of Chapter 4 are also necessary because of an erroneous statement in Buja (1984) which significantly reduces its applicability. This is proven in Subsection 4.1.2 which contains a counterexample and discusses the consequences.

After that, Section 4.2 follows the lines of Buja (1984) but replaces the concept of upper expectations by the concept of coherent upper previsions. As stated above, the results of Chapter 2 show that the latter concept is more appropriate in this decision theoretic setup. Accordingly, it is shown that the same result as in Buja (1984) is possible without any additional assumption on the involved (coherent) upper previsions. Here, the use of the extended decision theoretic concepts developed in Section 3.3 is crucial.

Thereafter, Section 4.3 turns over to hypothesis testing as a special case of decision theory. The existence of least favorable pairs in case of general coherent upper previsions follows from the previous results and, therefore, we provide an independent proof of an old result which has already been proven by Baumann (1968) in an equivalent setup.

Though the method of natural extension is one of the "three key ideas" [13] of the whole theory of imprecise probabilities due to P. Walley, its use still needs some justification especially for decision theoretic purposes. In particular, using this method raises two questions which are investigated in Chapter 5 and which have not been considered so far. In order to get an idea of these questions, note that coherent upper previsions are real functions $\overline{P} : \mathcal{K} \to \mathbb{R}$ where $\mathcal{K}$ is a set of bounded, measurable functions $f : \mathcal{X} \to \mathbb{R}$ on a sample space $(\mathcal{X}, \mathcal{A})$. The method of natural extension makes it possible to always extend coherent upper previsions on larger domains.

The first question, addressed in Section 5.2, is concerned with extensions from $\mathcal{K}$ to the whole space $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ consisting of all bounded, measurable functions $f : \mathcal{X} \to \mathbb{R}$. As stated before, coherent upper previsions should be used instead of precise probability measures because it is far more realistic to give an upper bound on the previsions/expectations/probabilities than to precisely specify these quantities in applications. However, the upper prevision $\overline{P}[f]$ is again a precise real number and this number $\overline{P}[f]$ will usually not be precisely known in real applications. So, a practitioner will hardly be able to decide if $\overline{P}$ is the "correct" upper prevision or if a slightly modified upper prevision $\overline{P}'$ is the correct one. Therefore, the important question arises what happens by an application of the methods of natural extension. Is the natural extension of $\overline{P}'$ still close to the natural extension of $\overline{P}$? The investigations in Section 5.2 show that, unfortunately, the answer to this question is not affirmative. Even more, arbitrarily small changes in $\overline{P}$ can have arbitrarily large effects on its natural extension in general and, therefore, applying natural extensions may lead to meaningless results. An example where this happens is given in Subsection 5.2.1. Fortunately, not all is lost. In Subsection 5.2.2, it is shown that it can be guaranteed in many situations that small changes in $\overline{P}$ on $\mathcal{K}$ only have

[13] see (Walley, 1991, p. 120–122)

small effects on the natural extension. However, these results are not fully satisfactory; hopefully, these initial investigations serve as a starting point for future research into this direction.

The second question is concerned with extensions of the sample space $(\mathcal{X},\mathcal{A})$ to a sample space $(\mathcal{X},\mathcal{A}')$ where $\mathcal{A}' \supset \mathcal{A}$: The method of natural extension enables to arbitrarily extend the algebra $\mathcal{A}$ of the initial sample space to that algebra which is most convenient. However, at least in decision theory and, especially in statistics, choosing $\mathcal{A}$ or the larger $\mathcal{A}'$ has a fundamental effect on the evaluations: the choice of the sample space determines the (randomized) decision functions and, in this way, extending the sample space leads to a larger set of valid (randomized) decision functions. Therefore, the important question arises if a (randomized) decision function which is optimal in the set of all (randomized) decision functions on $(\mathcal{X},\mathcal{A})$ is still optimal in the larger set of all (randomized) decision functions on $(\mathcal{X},\mathcal{A}')$ after natural extension. That is: Does optimality get lost by an application of natural extension? If the answer was affirmative, one should always choose the whole power set of $\mathcal{X}$ for the sample space – but the power set may be too large to be handled successfully. Especially in case of $\mathcal{X} = \mathbb{R}$, this would be very cumbersome. Fortunately, such an approach is not necessary because it is proven in Section 5.3 that – after applying natural extension – there is no (randomized) decision function on $(\mathcal{X},\mathcal{A}')$ which is better than the best (randomized) decision function on $(\mathcal{X},\mathcal{A})$. The proof is rather involved and heavily relies on the results of Chapter 3 and Chapter 4.

Chapter 5 closes with Section 5.4 which is concerned with discretizing – a topic which increasingly attracts attention within the theory of imprecise probabilities; cf. Obermeier and Augustin (2007) and Troffaes (2008). As an application of the results from Section 5.2 and Section 5.3, it is shown in Section 5.4 that decision problems can approximately be solved by solving appropriate discretized decision problems.

As stated above, the decision theoretic investigations of the present book are especially done in view of applications in statistics. Accordingly, the final chapter is devoted to such an application namely estimating. While hypothesis testing under imprecise probabilities has been extensively studied – especially by T. Augustin in (Augustin, 1998) and (Augustin, 2002) on base of the Huber-Strassen theory[14], estimating has hardly been considered explicitly within the theory of coherent upper previsions so far. Chapter 6 develops a minimum distance estimator which is based on the following simple idea: Analogously to classical statistics, the data $x_1, \ldots, x_n$ are assumed to be independent identically distributed according to a coherent upper prevision $\overline{P}'_{\theta_0}$ where $\theta_0 \in \Theta$ is an unknown parameter which has to be estimated. Then, the data are used to build the empirical measure

$$\mathbb{P}^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

and the minimum distance estimator is that $\hat{\theta} \in \Theta$ such that $\mathbb{P}^{(n)}$ lies next to $\overline{P}'_{\hat{\theta}}$.

Due to the present state of research, Chapter 6 cannot be restricted to the sole investigation of the proposed minimum distance estimator but also has to develop some fundamentals of (frequentist) estimating under coherent upper previsions at first. This is necessary the more so as the minimum distance estimator is associated with the empirical

[14] see also Augustin (2002) for a review of the work following Huber and Strassen (1973)

process (which needs a somewhat more elaborated setting) and is justified by asymptotic arguments (but an elaborated asymptotic theory of imprecise probabilities is still missing).

These preparations are done in Sections 6.2 and 6.3. In Section 6.4 it is proven that the distance between the empirical measure and the correct imprecise probability converges to zero and that the minimum distance estimator is consistent. Next, Section 6.5 is concerned with the calculation of the estimator. It is shown that the distances can be approximately calculated by use of the discretization method developed in Section 5.4. After discretizing, the (approximate) distance between the empirical measure and the coherent upper prevision can be calculated by solving a linear program. This linear program only modestly increases with the number of observations so that the minimum distance estimator can also be calculated for many observations. Finally, the estimator is applied in three simulation studies which indicate that it can indeed be applied and may lead to good results in real applications. The estimator has been programmed in R and is already publicly available as R package "imprProbEst"; cf. Hable (2008a).

Chapter 2

Topological aspects of imprecise probabilities

2.1 Introduction

Recently, the *Society for Imprecise Probability Theory and Applications (SIPTA)* has changed its name and is now called *Society for Imprecise Probability: Theories and Applications (SIPTA)* in order "to emphasize that there are theories of imprecise probability, rather than a single theory" [1]. Indeed, a short glance at de Cooman et al. (2007) shows that there a even many theories of imprecise probabilities. Two of the most important ones are the theory of coherent upper previsions (or coherent lower previsions) developed by P. Walley (see Walley (1991)) and the theory of F-probabilities developed by K. Weichselberger (see Weichselberger (2000) and Weichselberger (2001)). A first superficial reading of Walley (1991) and Weichselberger (2001) gives the impression that both concepts were most different from each other even though they are not. This totally wrong impression is the unfortunate consequence of the most different presentations used in these books. While Weichselberger (2001) is written in terms of random variables, probability measures and expectations (i.e. in terms of classical probability theory), Walley (1991) is written in terms of gambles, buying prices and previsions. As a consequence, many statisticians will highly underestimate the importance of Walley (1991) within the mathematical theory of statistics.

Since most part of the work presented in the following is motivated by applications in statistics, the presentation is based on concepts which are fundamental in traditional probability theory – such as sample spaces and probability measures – or which are close to traditional concepts – such as probability charges. Probability charges are similar to probability measures, the only difference is that σ-additivity is relaxes to finite additivity in the definition of probability charges. Accordingly, probability charges are often called "finitely additive probability measures" (e.g. in Dunford and Schwartz (1958)); the term "probability charge" originates from Bhaskara Rao and Bhaskara Rao (1983).

The definition of probability charges and some basic facts about them are recalled in Section 2.2. Next, Section 2.3 essentially recalls the definition of coherent upper previsions due to P. Walley. However, as mentioned above, the presentation totally differs because, here, Walley's definitions and some basic results are reformulated in terms of sample

[1] (de Cooman et al., 2007, p. xi)

spaces and probability charges[2]. By doing this, some of the basics results are slightly generalized.

On the one hand, such a reformulation may increase the accessibility of Walley's theory within mathematical statistics – the formulation in terms of gambles and buying prices seems to significantly handicap a wider acceptance of the whole theory. On the other hand, this makes a comparision with Weichselberger's theory more straightforward.

Section 2.3 investigates the concept of "upper expectations" which has been defined by Buja (1984) within robust statistics. Though this relatively old concept has been disregarded within the theory of imprecise probabilities so far, it may play an interesting part as a link because it lies somewhere between coherent upper previsions and F-probabilities. By use of this intermediate step, a comparison between Walley's and Weichselberger's concepts gets more easier: The only difference between coherent upper previsions and upper expectations is the fact that the latter concept insists on σ-additivity while Walley's concept of coherent upper previsions dispenses with σ-additivity. So, in a sense, upper expectations turn out to be a special case of coherent upper prevision. Next, F-probabilities can be defined as special upper expectations.

Most part of Section 2.3 is concerned with topological properties of upper expectations. This is because many theoretical evaluations in decision theory make use of minimax theorems and these minimax theorems heavily rely on topology. With this objective in mind, it follows from the investigations of Section 2.3 that the topological properties of coherent upper previsions are more convenient than the ones of upper expectations. This is the reason why the remaining parts of this book mainly deal with the concept of coherent upper previsions.

Though there are some differences between coherent upper prevision and upper expectations, the results from Section 2.5 shows that this is no longer true from an abstract point of view. In Section 2.5, it is proven that every coherent upper prevision can be represented by an upper expectation by use of the Stone representation theorem. Therefore, both concepts may be considered as equivalent. This is not only theoretically interesting but also beneficial later on: By use of this representation, concepts which rely on σ-additivity can also be defined for coherent upper prevision in a canonical way even though σ-additivity is usually not available in case of coherent upper previsions; see Subsection 3.3.3.

2.2 Precise probabilities

Before different concepts of imprecise probabilities are introduced, some notation has to be fixed and some facts about precise probabilities have to be recalled:

2.2.1 Probability charges

Let Ω be a set and $\mathcal{A}$ an algebra on Ω.

Definition 2.1 *A* bounded charge on $(\Omega, \mathcal{A})$ *is a map*

$$\mu : \ \mathcal{A} \ \to \ \mathbb{R}$$

such that

$$\mu(\emptyset) \ = \ 0 \tag{2.1}$$

[2] Reformulations in terms of probability charges have also been done by other authors, e.g. Troffaes (2008).

$$-\infty \;<\; \mu(A) \;<\; \infty \qquad \forall\, A \in \mathcal{A} \tag{2.2}$$

and

$$A_1, \ldots, A_n \in \mathcal{A},\;\; A_i \cap A_j = \emptyset \;\; (i \neq j) \quad \Rightarrow \quad \mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mu(A_n) \tag{2.3}$$

Bounded charges are also called *finitely additive, bounded, signed measures.* The term "charge" or "bounded charge" originates from Bhaskara Rao and Bhaskara Rao (1983). According to (Dunford and Schwartz, 1958, p. 240), the set of all bounded charges is denoted by $\mathrm{ba}(\Omega, \mathcal{A})$.

Definition 2.2 *A* probability charge on $(\Omega, \mathcal{A})$ *is a bounded charge* $P \in \mathrm{ba}(\Omega, \mathcal{A})$ *such that*

$$P(\Omega) \;=\; 1 \tag{2.4}$$

and

$$0 \;\leq\; P(A) \qquad \forall\, A \in \mathcal{A} \tag{2.5}$$

Probability charges are also called *finitely additive probability measures.* The term "probability charge" originates from Bhaskara Rao and Bhaskara Rao (1983) again.

In the following, $\mathrm{ba}(\Omega, \mathcal{A})$ is provided with some additional structures – namely with a norm which makes it a Banach space, with an ordering which makes it an L-space and with a weak topology. To this end, $\mathrm{ba}(\Omega, \mathcal{A})$ is identified with a suitable dual space at first.

I_A denotes the indicator function of A for every $A \subset \Omega$. Functions $s : \Omega \to \mathbb{R}$ of form

$$s: \;\; \omega \;\to\; \sum_{k=1}^{m} a_k I_{A_k}(\omega)\,, \qquad a_1, \ldots, a_m \,\in\, \mathbb{R}\,,\;\; A_1, \ldots, A_m \;\in\; \mathcal{A}\,, \quad m \in \mathbb{N}$$

are called $\mathcal{A}$-simple functions.
Let $\mathcal{L}_\infty(\Omega)$ denote the vector space of all functions $f : \Omega \to \mathbb{R}$ such that $\sup_{\omega\in\Omega} |f(\omega)| < \infty$. Provided with the norm

$$\|f\| \;=\; \sup_{\omega\in\Omega} |f(\omega)|$$

$\mathcal{L}_\infty(\Omega)$ is a Banach space, which contains every $\mathcal{A}$-simple function s. Let $\mathcal{L}_\infty(\Omega, \mathcal{A})$ denote the norm-closure of the set of all $\mathcal{A}$-simple functions in $\mathcal{L}_\infty(\Omega)$

$$\mathcal{L}_\infty(\Omega, \mathcal{A}) \;=\; c\ell_{\|\cdot\|}\big(\{s : \Omega \to \mathbb{R} \,|\; s \text{ is an } \mathcal{A}\text{-simple function}\}\big) \tag{2.6}$$

Hence, $\mathcal{L}_\infty(\Omega, \mathcal{A})$ is also a Banach space. If $\mathcal{A}$ is even a σ-algebra, this definition of $\mathcal{L}_\infty(\Omega, \mathcal{A})$ coincides with the more common definition in case of σ-algebras; cf. (Dunford and Schwartz, 1958, p. 240):

Lemma 2.3 *If* $\mathcal{A}$ *is a* σ*-algebra on* Ω*, then*

$$\mathcal{L}_\infty(\Omega, \mathcal{A}) \;=\; \big\{f : \Omega \to \mathbb{R} \;\big|\; f \textit{ is bounded and } \mathcal{A}\textit{-measurable}\big\}$$

Note that $\mathcal{L}_\infty(\Omega,\mathcal{A}) = \mathcal{L}_\infty(\Omega)$ if $\mathcal{A}$ is the power set of Ω.

For every $\mu \in \mathrm{ba}(\Omega,\mathcal{A})$ and every $f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$, the integral $\int f\,d\mu$ is defined in the following way (cf. e.g. (Dunford and Schwartz, 1958, § III.2)):

$$\int s\,d\mu \;=\; \sum_{k=1}^{m} a_k \mu(A_k) \tag{2.7}$$

for every $\mathcal{A}$ - simple function $s = \sum_{k=1}^{m} a_k I_{A_k}$.

$$\int f\,d\mu \;=\; \lim_{n} \int s_n\,d\mu \tag{2.8}$$

where $(s_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{A}$ - simple functions such that $\|s_n - f\| \xrightarrow[n\to\infty]{} 0$.

The following lemma says that the bounded charges $\mu \in \mathrm{ba}(\Omega,\mathcal{A})$ can be identified with the continuous linear functionals on $\mathcal{L}_\infty(\Omega,\mathcal{A})$.

Theorem 2.4 *For every* $\mu \in \mathrm{ba}(\Omega,\mathcal{A})$,

$$\mathcal{L}_\infty(\Omega,\mathcal{A}) \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; \int f\,d\mu$$

is a continuous linear functionals on $\mathcal{L}_\infty(\Omega,\mathcal{A})$.
Conversely, for every continuous linear functional

$$T:\; \mathcal{L}_\infty(\Omega,\mathcal{A}) \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; T(f)$$

on $\mathcal{L}_\infty(\Omega,\mathcal{A})$*, there is a unique* $\mu \in \mathrm{ba}(\Omega,\mathcal{A})$ *such that*

$$T(f) \;=\; \int f\,d\mu \qquad f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$$

That is, $\mathrm{ba}(\Omega,\mathcal{A})$ *is the dual space of* $\mathcal{L}_\infty(\Omega,\mathcal{A})$.
Confer (Dunford and Schwartz, 1958, Theorem IV.5.1).

Notation 2.5 *Since the elements of* $\mathrm{ba}(\Omega,\mathcal{A})$ *are rather considered as linear functionals on* $\mathcal{L}_\infty(\Omega,\mathcal{A})$ *than as set functions on* $\mathcal{A}$*, the notation*

$$\mu[f] \;=\; \int f\,d\mu\,, \qquad \mu \in \mathrm{ba}(\Omega,\mathcal{A})\,, \quad f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$$

is used in the following.

Norm:

As dual space of the Banach space $\mathcal{L}_\infty(\Omega,\mathcal{A})$, $\mathrm{ba}(\Omega,\mathcal{A})$ is again a Banach space with norm

$$\|\mu\| \;=\; \sup\big\{\mu[f] \;\big|\; \|f\|_\infty \le 1\big\}$$

This norm coincides with the total variation norm on $\mathrm{ba}(\Omega,\mathcal{A})$; cf. (Dunford and Schwartz, 1958, Theorem IV.5.1).

Ordering:

As dual space of $\mathcal{L}_\infty(\Omega,\mathcal{A})$, $\mathrm{ba}(\Omega,\mathcal{A})$ has a natural partial ordering $\le$

$$\mu_1 \;\le\; \mu_2 \qquad \Leftrightarrow \qquad \mu_1[f] \;\le\; \mu_2[f] \quad \forall\, f \ge 0\,,\;\; f \;\in\; \mathcal{L}_\infty(\Omega,\mathcal{A}) \tag{2.9}$$

As usual, $\mu_1 \le \mu_2$ is also denoted by $\mu_2 \ge \mu_1$. Note, that it is only required that $\mu_1[f] \le \mu_2[f]$ for every *non-negative* function f. In case of $\mu_1[f] \le \mu_2[f]$ for every function f, linearity implies $\mu_1 = \mu_2$.

Theorem 2.6 *ba*$(\Omega, \mathcal{A})$ *is an L-space.*
(Cf. (Bhaskara Rao and Bhaskara Rao, 1983, Theorem 2.2.1).)

Especially, $\text{ba}(\Omega, \mathcal{A})$ is a Dedekind complete Banach lattice. All these basic notions from lattice theory are collocated in Subsection 8.1.
The positive elements of $\text{ba}(\Omega, \mathcal{A})$ are the elements $\mu \in \text{ba}(\Omega, \mathcal{A})$ such that $\mu \geq 0$. It is easy to see that

$$\|\mu\| \;=\; \mu\big[I_\Omega\big] \qquad \forall\, \mu \,\geq\, 0\,, \quad \mu \in \text{ba}(\Omega, \mathcal{A}) \tag{2.10}$$

Furthermore,

$$\mu \;\geq\; 0 \qquad \Leftrightarrow \qquad \mu(A) \;\geq\; 0 \quad \forall\, A \in \mathcal{A} \tag{2.11}$$

That is, the usual order on bounded charges or bounded, signed measures coincides with the order defined by (2.9).[3] Therefore, the probability charges are precisely the normed ($\|\mu\| = 1$), positive elements of the L-space $\text{ba}(\Omega, \mathcal{A})$. Accordingly, the set of all probability charges is denoted by $\text{ba}_1^+(\Omega, \mathcal{A})$.

Furthermore, the probability charges are precisely the *precise coherent previsions on* $\mathcal{L}_\infty(\Omega, \mathcal{A})$ – a term which is common in the theory of imprecise previsions according to Walley (1991).

Weak*-topology:

$\text{ba}(\Omega, \mathcal{A})$ can be endowed with the weak*-topology. This is the weakest topology such that, for every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$,

$$\Lambda_f \;:\; \text{ba}(\Omega, \mathcal{A}) \;\to\; \mathbb{R}\,, \qquad \mu \;\mapsto\; \Lambda_f(\mu) \;=\; \mu[f]$$

is continuous. Some facts about this topology are collocated in the Appendix.
Later on, additional topologies of the same kind will be introduced. To unify notation, the weak*-topology is called $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - *topology* on $\text{ba}(\Omega, \mathcal{A})$ – as in (Dunford and Schwartz, 1958, p. 420).
Now, $\text{ba}(\Omega, \mathcal{A})$ already has two different topologies, namely the norm-topology and the (weaker) $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - topology. To make clear what topology is used, topological terms such as compact, open, closure etc. usually are denoted by norm-compact, $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - open, norm-closure etc.

The following theorem is one reason why the $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - topology is very useful.

Theorem 2.7 *The set of all probability charges* $\text{ba}_1^+(\Omega, \mathcal{A})$ *is* $\mathcal{L}_\infty(\Omega, \mathcal{A})$ *- compact in* $\text{ba}(\Omega, \mathcal{A})$.

Proof: $\text{ba}(\Omega, \mathcal{A})$ is the dual space of $\mathcal{L}_\infty(\Omega, \mathcal{A})$ (cf. Theorem 2.4). Hence, the closed unit sphere $\{\mu \in \text{ba}(\Omega, \mathcal{A}) \,|\, \|\mu\| \leq 1\}$ is $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - compact in $\text{ba}(\Omega, \mathcal{A})$ according to (Dunford and Schwartz, 1958, Theorem V.4.2).

Because of $\text{ba}_1^+(\Omega, \mathcal{A}) \;\subset\; \text{ba}(\Omega, \mathcal{A}) \,|\, \|\mu\| \leq 1\}$, it is enough to show that $\text{ba}_1^+(\Omega, \mathcal{A})$ is $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - closed. The latter statement is an easy consequence of Theorem 8.24 b). □

[3] (2.11) may be proven in the following way: Let $(s_n)_{n\in\mathbb{N}}$ be a sequence of simple functions such that $\|s_n - h\| \to 0$ for any $h \geq 0$. Then, $\hat{s}_n(\omega) = \max\{s_n(\omega)\,,\, 0\}$ defines a sequence of non-negative functions such that $\|\hat{s}_n - h\| \to 0$ and $\mu[h] = \lim_n \mu[\hat{s}_n] \geq 0$.

2.2.2 Probability measures

Let Ω be a set and $\mathcal{A}$ a σ-algebra on Ω.

Definition 2.8 *A* bounded, signed measure on $(\Omega, \mathcal{A})$ *is a map*

$$\mu : \ \mathcal{A} \ \to \ \mathbb{R}$$

such that

$$\mu(\emptyset) \ = \ 0 \tag{2.12}$$

$$-\infty \ < \ \mu(A) \ < \ \infty \qquad \forall\, A \in \mathcal{A} \tag{2.13}$$

and $\big(A_n\big)_{n\in\mathbb{N}} \subset \mathcal{A},\ \ A_n \cap A_m = \emptyset \ \ (n \neq m)$ *implies*

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n) \tag{2.14}$$

According to (Dunford and Schwartz, 1958, p. 240), the set of all bounded, signed measures is denoted by $\mathrm{ca}(\Omega, \mathcal{A})$. Property (2.14) is called σ-additivity. This property is the only difference to the definition of bounded charges, where σ-additivity is relaxed to finite additivity (2.3). Since

$$\mathrm{ca}(\Omega, \mathcal{A}) \ \subset \ \mathrm{ba}(\Omega, \mathcal{A})\,,$$

$\mathrm{ca}(\Omega, \mathcal{A})$ inherits the definition of the integral $\mu[f] = \int f\, d\mu$, the norm $\|\cdot\|$ and the ordering $\leq$ from $\mathrm{ba}(\Omega, \mathcal{A})$.

Theorem 2.9 $\mathrm{ca}(\Omega, \mathcal{A})$ *is a band in the L-space* $\mathrm{ba}(\Omega, \mathcal{A})$*. Hence,* $\mathrm{ca}(\Omega, \mathcal{A})$ *is itself an L-space.*
(Cf. (Bhaskara Rao and Bhaskara Rao, 1983, Theorem 2.4.2).)

Definition 2.10 *A* probability measure on $(\Omega, \mathcal{A})$ *is a bounded, signed measure* $P \in \mathrm{ba}(\Omega, \mathcal{A})$ *such that*

$$P(\Omega) \ = \ 1 \tag{2.15}$$

and

$$0 \ \leq \ P(A) \qquad \forall\, A \in \mathcal{A} \tag{2.16}$$

The probability measures are precisely the normed ($\|\mu\| = 1$), positive elements of the L-space $\mathrm{ca}(\Omega, \mathcal{A})$. Therefore, the set of all probability measures is denoted by $\mathrm{ca}_1^+(\Omega, \mathcal{A})$.

Furthermore, the probability measures are precisely the *precise coherent previsions* $P : \mathcal{L}_\infty(\Omega, \mathcal{A}) \to \mathbb{R}$ which are *σ-smooth*:

$$f_n(\omega) \searrow f(\omega) \quad \forall\, \omega \in \Omega \qquad \Rightarrow \qquad \lim_{n\to\infty} P[f_n] \ = \ P[f]$$

Theorem 2.11

a) *ca*$(\Omega, \mathcal{A})$ *is* $\mathcal{L}_\infty(\Omega, \mathcal{A})$ *- dense in ba*$(\Omega, \mathcal{A})$.

b) *$\mathrm{ba}_1^+(\Omega,\mathcal{A})$ is the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure of $\mathrm{ca}_1^+(\Omega,\mathcal{A})$ in $\mathrm{ba}(\Omega,\mathcal{A})$.*

Proof: Let $c\ell$ denote the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure in $\mathrm{ba}(\Omega,\mathcal{A})$. Then, convexity of $\mathrm{ca}(\Omega,\mathcal{A})$ and $\mathrm{ca}_1^+(\Omega,\mathcal{A})$ imply that

$$c\ell\big(\mathrm{ca}(\Omega,\mathcal{A})\big) \qquad \text{and} \qquad c\ell\big(\mathrm{ca}_1^+(\Omega,\mathcal{A})\big)$$

are the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closed convex hulls respectively (cf. (Dunford and Schwartz, 1958, Theorem V.2.1)).

a) Note, that $\sup_{\mu\in\mathrm{ca}(\Omega,\mathcal{A})}\mu[f] = \sup_{\mu\in\mathrm{ba}(\Omega,\mathcal{A})}\mu[f]$ for every $f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$. Hence, part a) follows from Theorem 8.26 where $\mathcal{V}=\mathrm{ca}(\Omega,\mathcal{A})$ and $M=\mathrm{ba}(\Omega,\mathcal{A})$.

b) Put $\overline{P}[f] := \sup_{P\in\mathrm{ca}_1^+(\Omega,\mathcal{A})} P[f] = \sup_{\omega\in\Omega} f(\omega)$ for every $f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$. Then, it is easy to see that

$$\big\{\mu\in\mathrm{ba}(\Omega,\mathcal{A}) \;\big|\; \mu[f]\le\overline{P}[f] \;\; \forall\, f\in\mathcal{L}_\infty(\Omega,\mathcal{A})\big\} \;=\; \mathrm{ba}_1^+(\Omega,\mathcal{A})$$

and b) follows from Theorem 8.26 where $\mathcal{V}=\mathrm{ca}_1^+(\Omega,\mathcal{A})$ and $M=\mathrm{ba}(\Omega,\mathcal{A})$ □

2.2.3 σ-additivity

Some readers may feel that allowing probabilities not to be σ-additive is illegal and that one should only use probability measures and leave out all other probability charges. Indeed, the use of probability charges which are not σ-additive is justified by Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung (1933)*, p. 14:

> "For infinite fields, on the other hand, the Axiom of Continuity[4], VI, proved to be independent of Axioms I - V. Since the new axiom is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning, as has been done, for example, in the case of Axioms I - V in § 2 of the first chapter. For, in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI*[5]. This limitation has been found expedient in researches of the most diverse sort."[6]

Consequentially, Le Cam (1986) argues on p. 2:

> "However, we need to point out that, when $\mathcal{A}$ is not finite, the traditional description makes use of certain mathematical constructs which have been introduced for convenience in other contexts. Since there the matter of mathematical convenience is the relevant one, one may feel free to vary the constructs as long as their relations to the real world are not affected."

The use of σ-additive probability charges (i.e. probability measures) is appropriate in many situations. The main disadvantage of dispensing with σ-additivity is the fact that the usual Radon-Nikodym theorem gets lost then (see (Bhaskara Rao and Bhaskara Rao, 1983, § 6.3)) so that we are not able to work with densities. However, in case of imprecise probabilities, it is nevertheless often advantageous to dispense with σ-aditivity as will be seen later on, in particular in Section 2.4 and Chapter 4.

[4] together with finite additivity, Axiom VI corresponds to σ-additivity

[5] together with finite additivity, Axiom VI corresponds to σ-additivity

[6] The English translation of the quote is taken from (Kolmogorov, 1956, p. 15).

2.3 Coherent upper previsions

This section recalls the definitions of coherent upper previsions (and coherent lower previsions) according to P. Walley although the presentation is quite different to that one in Walley (1991) and some basic results are slightly generalized.

Let Ω be a set and $\mathcal{A}$ an algebra on Ω. Let $\mathcal{K}$ be any subset of $\mathcal{L}_\infty(\Omega,\mathcal{A})$.

Definition 2.12 *Let $\overline{P}$ be a map*

$$\mathcal{K} \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; \overline{P}[f]$$

and put $\mathcal{M} \;:=\; \big\{P \in \mathrm{ba}_1^+(\Omega,\mathcal{A}) \;\big|\; P[f] \le \overline{P}[f] \;\;\forall\, f \in \mathcal{K}\big\}$. $\overline{P}$ *is called* coherent upper prevision on $\mathcal{K}$ *if*

- $\mathcal{M} \;\ne\; \emptyset$
- $\sup\limits_{P \in \mathcal{M}} P[f] \;=\; \overline{P}[f]$

$\mathcal{M}$ *is called the* credal set of $\overline{P}$ (on $(\Omega,\mathcal{A})$).
If $\overline{P}$ is a coherent upper prevision on $\mathcal{K}$,

$$\underline{P}:\; -\mathcal{K} \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; -\overline{P}[-f]$$

is called coherent lower prevision on $-\mathcal{K}$.

The following proposition shows that this definition does not depend on $\mathcal{A}$ and describes how a coherent upper prevision can be extended to a coherent upper prevision on any set $\mathcal{K}'$ such that $\mathcal{K} \subset \mathcal{K}' \subset \mathcal{L}_\infty(\Omega)$. As a consequence, it can always be assumed that coherent upper previsions are defined on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ or even on $\mathcal{L}_\infty(\Omega)$.

Proposition 2.13 *Let $\mathcal{A}'$ be another algebra on Ω such that $\mathcal{A} \subset \mathcal{A}'$. Let $\overline{P}:\; \mathcal{K} \to \mathbb{R}$ be a coherent upper prevision on $\mathcal{K}$ and let $\mathcal{M}$ be its credal set on $(\Omega,\mathcal{A})$. Furthermore, let $\mathcal{K}'$ be a set of functions such that $\mathcal{K} \subset \mathcal{K}' \subset \mathcal{L}_\infty(\Omega)$. Put*

$$\mathcal{M}' \;:=\; \big\{P' \in \mathrm{ba}_1^+(\Omega,\mathcal{A}') \;\big|\; P'[f'] \le \overline{P}[f'] \;\;\forall\, f' \in \mathcal{K}\big\}$$

Then,

- $\mathcal{M}' \;\ne\; \emptyset$
- $\sup\limits_{P' \in \mathcal{M}'} P'[f] \;=\; \overline{P}[f] \quad \forall\, f \in \mathcal{K}$

$\mathcal{M}'$ *is called the* credal set of $\overline{P}$ on $(\Omega,\mathcal{A}')$. *Furthermore,*

$$\overline{P}'[f'] \;=\; \sup_{P' \in \mathcal{M}'} P'[f']\,, \qquad f' \in \mathcal{K}'$$

defines a coherent upper prevision on $\mathcal{K}'$ which is an extension of $\overline{P}$. $\mathcal{M}'$ is the credal set of $\overline{P}'$ on $(\Omega,\mathcal{A}')$.

Proof: $S : \mathcal{L}_\infty(\Omega,\mathcal{A}') \to \mathbb{R}$ defined by $S(f') = \sup_{\omega\in\Omega} f'(\omega)$ is a sublinear functional. Take any $P \in \mathcal{M}$. Then, $P[f] \le S(f) \quad \forall\, f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$ and, according to (Dunford and Schwartz, 1958, Theorem II.3.10), P can be extended to a linear functional P' on $\mathcal{L}_\infty(\Omega,\mathcal{A}')$ such that

$$P'[f'] \;\le\; S(f') \;=\; \sup_{\omega\in\Omega} f'(\omega) \qquad \forall\, f' \in \mathcal{L}_\infty(\Omega,\mathcal{A}')$$

$|P'[f']| \le \|f'\|$ implies that P' is a (norm-)continuous linear functional and, therefore, $P \in \mathrm{ba}(\Omega,\mathcal{A}')$. From

$$P'[f'] \;=\; -P'[-f'] \;\ge\; -S(-f') \;\ge\; 0 \qquad \forall\, f' \;\ge\; 0\,, \quad f' \in \mathcal{L}_\infty(\Omega,\mathcal{A}')$$

and

$$1 \;=\; -\sup_{\omega\in\Omega}\big(-I_\Omega(\omega)\big) \;\le\; -P'\big[-I_\Omega\big] \;=\; P'\big[I_\Omega\big] \;\le\; \sup_{\omega\in\Omega} I_\Omega(\omega) \;=\; 1$$

it follows that $P' \in \mathrm{ba}_1^+(\Omega,\mathcal{A}')$.

Hence, $\mathcal{M}' \ne \emptyset$ and $\sup\limits_{P'\in\mathcal{M}'} P'[f] \;=\; \overline{P}[f] \quad \forall\, f \in \mathcal{K}$.

The remaining statements of the proposition follow directly from the definitions. □

Especially, Proposition 2.13 shows that the definitions of coherent upper/lower previsions are only reformulations of the original definitions (Walley, 1991, § 2.5.1); cf. also (Walley, 1991, § 3.3.3). The extension procedure described in Proposition 2.13 is called *natural extension* in (Walley, 1991, § 3).

It does not make any essential difference if coherent *upper* or coherent *lower* previsions are considered. In this book, most of the results are formulated in terms of coherent *upper* previsions because the main purpose are decision theoretic investigations with applications in statistics. Here, the use of loss functions leads to upper risks which are defined via coherent upper previsions. If utility functions were considered instead (which is quite unusual in statistics), the results would accordingly have to be formulated in terms of coherent *lower* previsions.

The following well-known theorem (cf. (Walley, 1991, Theorem 2.5.5)) provides a characterization of coherent upper previsions.

Theorem 2.14 *$\overline{P}$ is a coherent upper prevision on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ if and only if*

- $\overline{P}[f] \;\le\; \sup\limits_{\omega\in\Omega} f(\omega) \qquad \forall\, f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$
- $\overline{P}[af] \;=\; a\overline{P}[f] \qquad \forall\, f \in \mathcal{L}_\infty(\Omega,\mathcal{A})\,,\ \forall\, a \in (0,\infty)$ *and*
- $\overline{P}[f_1 + f_2] \;\le\; \overline{P}[f_1] + \overline{P}[f_2] \qquad \forall\, f_1,\ f_2 \;\in\; \mathcal{L}_\infty(\Omega,\mathcal{A})$

Proposition 2.15 describes a common way for generating coherent upper previsions:

Proposition 2.15 *Let $\mathcal{V} \;\subset\; \mathrm{ba}_1^+(\Omega,\mathcal{A})$ be any non-empty subset of probability charges on $(\Omega,\mathcal{A})$. Then,*

$$\overline{P}[f] \;=\; \sup_{P\in\mathcal{V}} P[f]\,, \qquad f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$$

defines a coherent upper prevision on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ and the convex $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure of $\mathcal{V}$

$$\mathcal{M} \;=\; \mathcal{C}\ell co\big(\mathcal{V}\big)$$

is the credal set of $\overline{P}$ on $(\Omega,\mathcal{A})$.

Proof: It follows directly from the definitions that $\overline{P}$ is a coherent upper prevision. Note that $\mathcal{V} \subset \mathrm{ba}_1^+(\Omega,\mathcal{A})$ and $\mathrm{ba}_1^+(\Omega,\mathcal{A})$ is a $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closed convex subset of $\mathrm{ba}(\Omega,\mathcal{A})$ (cf. Theorem 2.7). Therefore,

$$c\ell\mathrm{co}\big(\mathcal{V}\big) \ \subset\ \mathrm{ba}_1^+(\Omega,\mathcal{A}) \tag{2.17}$$

Let $\mathcal{M}$ be the credal set of $\overline{P}$. Then, Theorem 8.26 implies

$$\begin{aligned}\mathcal{M} \ &=\ \Big\{P \in \mathrm{ba}(\Omega,\mathcal{A})\ \Big|\ P[f] \le \overline{P}[f]\quad \forall f \in \mathcal{L}_\infty(\Omega,\mathcal{A})\Big\} \cap \mathrm{ba}_1^+(\Omega,\mathcal{A}) \ = \\ &=\ c\ell\mathrm{co}\big(\mathcal{V}\big) \cap \mathrm{ba}_1^+(\Omega,\mathcal{A}) \ \overset{(2.17)}{=}\ c\ell\mathrm{co}\big(\mathcal{V}\big)\end{aligned}$$

□

Whereas Theorem 2.14 characterizes coherent upper previsions, the following corollary characterizes the credal sets. The corollary is a slight generalization of (Walley, 1991, Theorem 3.6.1), it says that there is a one-to-one correspondence between coherent upper previsions and $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - compact convex subsets of $\mathrm{ba}_1^+(\Omega,\mathcal{A})$.

Corollary 2.16 *A subset $\mathcal{M} \subset \mathrm{ba}_1^+(\Omega,\mathcal{A})$ is a credal set of a coherent upper prevision if and only if it is $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - compact and convex.*

Proof: Let $\mathcal{M}$ be $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - compact and convex. Put $\mathcal{V} = \mathcal{M}$ and define a coherent upper prevision $\overline{P}$ as in Proposition 2.15. Then, Proposition 2.15 implies that $\mathcal{M} = c\ell\mathrm{co}\big(\mathcal{M}\big)$ is the credal set of $\overline{P}$.

Conversely, let $\mathcal{M}$ be the credal set of some coherent upper prevision $\overline{P}$. Put $\mathcal{V} = \mathcal{M}$; the coherent upper prevision defined by $\mathcal{V} = \mathcal{M}$ as in Proposition 2.15 is again $\overline{P}$. By assumption, $\mathcal{M}$ is the credal set of $\overline{P}$ so that Proposition 2.15 implies

$$\mathcal{M} \ =\ c\ell\mathrm{co}\big(\mathcal{M}\big)$$

Hence, $\mathcal{M}$ is $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closed and convex. Since $\mathcal{M} \ \subset\ \mathrm{ba}_1^+(\Omega,\mathcal{A})$ and $\mathrm{ba}_1^+(\Omega,\mathcal{A})$ is $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - compact (cf. Theorem 2.7), it follows that $\mathcal{M}$ is $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - compact, too. □

2.4 Upper/lower expectations and F-probabilities

2.4.1 Definitions and basic properties

Within the concept of coherent upper previsions developed in Walley (1991), σ-additivity is mainly ignored. However, there is also the concept of upper expectations and F-probabilities developed in Buja (1984) and Weichselberger (2001) which insists on σ-additivity. Essentially, this means that only those coherent upper previsions $\overline{P}$ are considered which have a representation by a set of probability measures:

$$\overline{P}[f] \ =\ \sup_{P \in \mathcal{P}} P[f] \qquad \forall f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$$

where $\mathcal{P} \subset \mathrm{ca}_1^+(\Omega,\mathcal{A})$.

Let Ω be a set and $\mathcal{A}$ a σ-algebra on Ω. Let $\mathcal{K}$ be any subset of $\mathcal{L}_\infty(\Omega,\mathcal{A})$.

Definition 2.17 *Let $\overline{P}$ be a map*

$$\overline{P}\;:\quad \mathcal{K}\;\to\;\mathbb{R}\,,\qquad f\;\mapsto\;\overline{P}[f]$$

and put $\mathcal{P}\;:=\;\big\{P\in\mathrm{ca}_1^+(\Omega,\mathcal{A})\;\big|\;\;P[f]\le\overline{P}[f]\;\;\forall\,f\in\mathcal{K}\big\}$. $\overline{P}$ *is called* upper expectation on $\mathcal{K}$ *if*

- $\mathcal{P}\;\neq\;\emptyset$
- $\sup\limits_{P\in\mathcal{P}} P[f]\;=\;\overline{P}[f]$

$\mathcal{P}$ *is called the* structure of $\overline{P}$*.*
If $\overline{P}$ *is an upper expectation on* $\mathcal{K}$*,*

$$\underline{P}:\;\;-\mathcal{K}\;\to\;\mathbb{R}\,,\qquad f\;\mapsto\;-\overline{P}[-f]$$

is called lower expectation on $-\mathcal{K}$ *.*

Definition 2.18 *Let $\overline{P}$ be an upper expectation on $\mathcal{K}$. If*

$$\mathcal{K}\subset\big\{I_A\;\big|\;\;A\in\mathcal{A}\big\}$$

$\overline{P}$ *is also called* upper F-probability *and* $\underline{P}$ *is also called* lower F-probability *.*

The definition of upper expectations originates from Buja (1984), the definition of F-probabilities originates from Weichselberger (2000) and Weichselberger (2001) though Weichselberger uses the terms “lower/upper interval-limit of the F-probability” instead of “lower/upper F-Probability”. The term “structure” also stems from Weichselberger (2000); originally, structures are defined in case of F-probabilities only. In Definition 2.17, the term “structure” is adopted for every upper expectation.

The following proposition describes how an upper expectation can be extended to an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. As a consequence, it can always be assumed that upper expectations are defined on the whole space $\mathcal{L}_\infty(\Omega,\mathcal{A})$.

Proposition 2.19 *Let* $\overline{P}:\;\mathcal{K}\to\mathbb{R}$ *be an upper expectation on* $\mathcal{K}\;\subset\;\mathcal{L}_\infty(\Omega,\mathcal{A})$ *and let* $\mathcal{P}$ *be its structure on* $(\Omega,\mathcal{A})$*. Then,* $\overline{P}$ *can be extended to an upper expectation on* $\mathcal{L}_\infty(\Omega,\mathcal{A})$ *by*

$$\overline{P}[f]\;:=\;\sup_{P\in\mathcal{P}}P[f]\,,\qquad f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$$

$\mathcal{P}$ *is also the structure of the extended upper expectation..*

Proof: This is a direct consequence of the definitions. □

Remark 2.20 *Proposition 2.19 is considerably weaker than its analog in case of coherent upper previsions (Proposition 2.13). This is due to insistence on σ-additivity for the elements of the structure: The proof of Proposition 2.13 (in case of coherent upper previsions) is based on the fact that every probability charge on $\mathcal{A}$ can be extended to a probability charge on any $\mathcal{A}'\supset\mathcal{A}$ according to the Hahn-Banach theorem (e.g. (Dunford and Schwartz, 1958, Theorem II.3.10)). However, there is no analog of the Hahn-Banach theorem if we insist on σ-additivity. It is possible that a probability measure on $\mathcal{A}$ can not be extended to a probability measure on some σ-algebra $\mathcal{A}'\supset\mathcal{A}$. Such problems does not* *only arise in artificial cases. For example, let $\overline{P}$ be the upper prevision whose structure only consists of the standard normal distribution $P := \mathcal{N}(0,1)$. That is $\overline{P} = P := \mathcal{N}(0,1)$, $\mathcal{A} = \mathbb{B}$. Assume that $\overline{P}$ admits an extension to an upper expectation on the power set $2^{\mathbb{R}}$ of $\mathbb{R}$. Then, the structure of the extended upper expectation consists of probability measures on $2^{\mathbb{R}}$ which are extensions of $P = \mathcal{N}(0,1)$. Let P' be such an extension of $P = \mathcal{N}(0,1)$, f be the density of $\mathcal{N}(0,1)$ with respect to the Lebesgue measure λ and put $g := 1/f$. Then, λ' defined by $\lambda'(A) = \int_A g\,dP'\ \forall A \in 2^{\mathbb{R}}$ is an extension of λ to $2^{\mathbb{R}}$. However, it is known that the existence of an extension λ' neither can be proven nor disproven. That is, the only way to get an extension of $\overline{P} = \mathcal{N}(0,1)$ on $2^{\mathbb{R}}$ is to introduce the existence of such an extension as a new axiom in mathematics; confer (Hoffmann-Jørgensen, 1994b, p. 513).*

The following theorem states that every upper expectation is a coherent upper prevision. Of course, the corresponding credal set does, in general, not coincide with the structure but there is a strong relationship between these sets: the corresponding credal set is the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure of the structure.
However, credal set and structure do coincide sometimes – this instance characterizes an important special case (cf. Theorem 2.27).

Proposition 2.21

a) *Let $\overline{P}$ be an upper expectation on $\mathcal{K} \subset \mathcal{L}_\infty(\Omega,\mathcal{A})$. Then, $\overline{P}$ is also a coherent upper prevision on $\mathcal{K}$.*

b) *Let $\overline{P}$ be an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ with structure $\mathcal{P}$. Then, $\overline{P}$ is also a coherent upper prevision on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ and its credal set $\mathcal{M}$ (on $(\Omega,\mathcal{A})$) is equal to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure of the structure $\mathcal{P}$*

$$\mathcal{M} \;=\; c\ell(\mathcal{P})$$

In particular, $\mathcal{P}$ is relatively compact with respect to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ba}(\Omega,\mathcal{A})$.

Proof:

a) Let $\mathcal{P}$ be the structure of the upper expectation $\overline{P}$ and put

$$\mathcal{M} \;:=\; \left\{ P \in \mathrm{ba}_1^+(\Omega,\mathcal{A}) \;\middle|\; P[f] \le \overline{P}[f] \;\; \forall f \in \mathcal{K} \right\}$$

Then, $\mathcal{P} \subset \mathcal{M}$ and, therefore, it follows that $\mathcal{M} \neq \emptyset$ and

$$\sup_{P \in \mathcal{M}} P[f] \;=\; \overline{P}[f] \quad \forall f \in \mathcal{K}$$

Hence, $\overline{P}$ is a coherent upper prevision.

b) Convexity of $\mathcal{P}$ implies that $c\ell(\mathcal{P})$ is the convex $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closure of $\mathcal{P}$; cf. (Dunford and Schwartz, 1958, Theorem V.2.1). Then, statement b) follows from Proposition 2.15 where $\mathcal{V} := \mathcal{P}$.
Furthermore, this implies that $\mathcal{P}$ is relatively compact in the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology; cf. Corollary 2.16.

□

Remark 2.22 *Assumption $\mathcal{K} = \mathcal{L}_\infty(\Omega,\mathcal{A})$ is crucial in Proposition 2.21 b). The mathematical reason is as follows: Let $\overline{P}$ be an upper expectation on $\mathcal{K} \subset \mathcal{L}_\infty(\Omega,\mathcal{A})$. Then, the natural extension (Proposition 2.13) for coherent upper previsions and the extension procedure for upper expectations (according to Proposition 2.19), in general, do not coincide on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ because the structure $\mathcal{P}$ and the credal set $\mathcal{M}$ usually do not coincide.*

Since the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ba}(\Omega,\mathcal{A})$ is useful in case of coherent upper previsions, it is suggesting to use the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ca}(\Omega,\mathcal{A})$ for upper expectations. This is the weakest topology on $\mathrm{ca}(\Omega,\mathcal{A})$ such that, for every $f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$,

$$\Lambda_f \,:\; \mathrm{ca}(\Omega,\mathcal{A}) \;\to\; \mathbb{R}\,, \qquad \mu \;\mapsto\; \Lambda_f(\mu) \;=\; \mu[f]$$

is continuous. Note, that this topology coincides with the subspace topology on $\mathrm{ca}(\Omega,\mathcal{A})$ generated by the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ba}(\Omega,\mathcal{A})$; cf. Lemma 8.25.

Proposition 2.23 describes a common way for generating upper expectations – it is the analog of Proposition 2.15.

Proposition 2.23 *Let $\mathcal{V} \subset \mathrm{ca}_1^+(\Omega,\mathcal{A})$ be any subset of probability measures on $(\Omega,\mathcal{A})$. Then,*

$$\overline{P}[f] \;=\; \sup_{P\in\mathcal{V}} P[f]\,, \qquad f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$$

defines an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. The structure $\mathcal{P}$ of $\overline{P}$ is given by the convex closure of $\mathcal{V}$ in $\mathrm{ca}(\Omega,\mathcal{A})$

$$\mathcal{P} \;=\; c\ell\mathrm{co}\big(\mathcal{V}\big)$$

where the term "closure" refers to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ca}(\Omega,\mathcal{A})$.
$\mathcal{V}$ is called prestructure of the upper expectation *$\overline{P}$.*

Proof: It is a direct consequence of the definitions that $\overline{P}$ is an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$.

All topological terms within this proof refer to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ca}(\Omega,\mathcal{A})$. It is an easy consequence of Theorem 8.24 b) that $\mathrm{ca}_1^+(\Omega,\mathcal{A})$ is closed in $\mathrm{ca}(\Omega,\mathcal{A})$. Therefore, convexity of $\mathrm{ca}_1^+(\Omega,\mathcal{A})$ implies

$$c\ell\mathrm{co}\,\big(\mathcal{V}\big) \;\subset\; \mathrm{ca}_1^+(\Omega,\mathcal{A}) \tag{2.18}$$

Let $\mathcal{P}$ be the credal set of $\overline{P}$. Then, Theorem 8.26 implies

$$\begin{aligned}\mathcal{P} \;&=\; \Big\{P \in \mathrm{ca}(\Omega,\mathcal{A}) \;\Big|\; P[f] \le \overline{P}[f] \;\;\forall f \in \mathcal{L}_\infty(\Omega,\mathcal{A})\Big\} \cap \mathrm{ca}_1^+(\Omega,\mathcal{A}) \;=\\ &=\; c\ell\mathrm{co}\,\big(\mathcal{V}\big) \cap \mathrm{ca}_1^+(\Omega,\mathcal{A}) \;\overset{(2.18)}{=}\; c\ell\mathrm{co}\,\big(\mathcal{V}\big)\end{aligned}$$

□

The term "prestructure" was originally defined in (Weichselberger, 2000, Defnition 2.5) in case of F-probabilities. Note that there is an important difference between the theory of F-probabilities and the theory of upper expectations:

Caution: *According to Proposition 2.23, a set $\mathcal{V} \subset \mathrm{ca}_1^+(\Omega,\mathcal{A})$ generates an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. Analogously, $\mathcal{V}$ may also generate an upper expectation $\overline{P}_{\mathcal{K}}$ on some*

$$\mathcal{K} \;\subset\; \mathcal{L}_\infty(\Omega,\mathcal{A})$$

by

$$\overline{P}_{\mathcal{K}}[f] \;=\; \sup_{P\in\mathcal{V}} P[f]\,, \qquad f\in\mathcal{K}$$

Of course, we have

$$\overline{P}_{\mathcal{K}}[f] \;=\; \overline{P}[f] \qquad \forall\, f\in\mathcal{K}$$

However, if we extend $\overline{P}_{\mathcal{K}}$ on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ according to Proposition 2.19 (the extension is again denoted by $\overline{P}_{\mathcal{K}}$), we usually have

$$\overline{P}_{\mathcal{K}}[f] \;\neq\; \overline{P}[f] \qquad \textit{for} \quad f\notin\mathcal{K}$$

because the structures of $\overline{P}$ and $\overline{P}_{\mathcal{K}}$ usually do not coincide.
Especially, this applies to F-probabilities, where $\mathcal{K}=\{I_A \mid A\in\mathcal{A}\}$. As a consequence, the structure of the F-probability $\overline{P}_{\mathcal{K}}$ generated by $\mathcal{V}$ does not coincide with the structure of the upper expectation $\overline{P}$ generated by $\mathcal{V}$ on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ in general!

The following corollary characterizes structures of upper expectations and establishes a one-to-one correspondence between upper expectations and $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closed convex subsets of $\mathrm{ca}(\Omega,\mathcal{A})$.

Corollary 2.24 *A subset $\mathcal{P}\subset\mathrm{ca}_1^+(\Omega,\mathcal{A})$ is a structure of an upper expectation if and only if it is $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - closed in $\mathrm{ca}(\Omega,\mathcal{A})$ and convex.*

Proof: All topological terms within this proof are with respect to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on $\mathrm{ca}(\Omega,\mathcal{A})$.

Let $\mathcal{P}$ be closed and convex. Put $\mathcal{V}=\mathcal{P}$ and define an upper expectation $\overline{P}$ as in Proposition 2.23. Then, Proposition 2.23 implies that $\mathcal{P}=c\ell\mathrm{co}\big(\mathcal{P}\big)$ is the structure of $\overline{P}$.

Conversely, let $\mathcal{P}$ be the structure of some upper expectation $\overline{P}$. Put $\mathcal{V}=\mathcal{P}$; the upper expectation defined by $\mathcal{V}=\mathcal{P}$ as in Proposition 2.23 is again $\overline{P}$. By assumption, $\mathcal{P}$ is the structure of $\overline{P}$ so that Proposition 2.23 implies

$$\mathcal{P} \;=\; c\ell\mathrm{co}\big(\mathcal{P}\big)$$

Hence, $\mathcal{P}$ is closed and convex. □

Corollary 2.24 is weaker than its analog in case of coherent upper previsions (Corollary 2.16): While credal sets are always compact (with respect to the considered topology), structures are not necessarily compact (with respect to the considered topology). However, compactness is an important property because it enables us to use minimax theorems throughout this book. Indeed some of the most important results in this book are based on minimax theorems.

Therefore, the next subsection is concerned with the investigation of necessary and sufficient conditions for compactness of structures.

2.4.2 Continuous upper expectations

Let Ω be a set and $\mathcal{A}$ a σ-algebra on Ω. In order to investigate necessary and sufficient conditions for compactness of structures, we have to introduce some terminology:

Definition 2.25 *An upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$ is called* continuous *if for every sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{L}_\infty(\Omega,\mathcal{A})$ and $f \in \mathcal{L}_\infty(\Omega,\mathcal{A})$*

$$f_n(\omega) \searrow f(\omega) \;\; \forall\, \omega \in \Omega \qquad \Rightarrow \qquad \lim_{n\to\infty} \overline{P}[f_n] \,=\, \overline{P}[f]$$

In case of F-probabilities, this definition originates from (Weichselberger, 2000, Definition 2.6).

Definition 2.26 *A set $\mathcal{V} \subset \mathrm{ca}_1^+(\Omega,\mathcal{A})$ of probability measures is called uniformly dominated by $\mu \in \mathrm{ca}(\Omega,\mathcal{A})$, $\mu \geq 0$, if*

$$\forall\,\varepsilon > 0 \quad \exists\,\delta > 0\,, \quad \textit{such that, for every } A \in \mathcal{A} \textit{ and } P \in \mathcal{V}\textit{:}$$

$$\mu(A) < \delta \;\;\Rightarrow\;\; P(A) < \varepsilon$$

Proposition 2.21 claims a strong connection between upper expectations and coherent upper previsions on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. It says that every upper expectation is also a coherent upper prevision. So, an upper expectation $\overline{P}$ corresponds to a structure $\mathcal{P} \subset \mathrm{ca}_1^+(\Omega,\mathcal{A})$ and to a credal set $\mathcal{M} \subset \mathrm{ba}_1^+(\Omega,\mathcal{A})$. According to Proposition 2.21, $\mathcal{M}$ is the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-closure of $\mathcal{P}$ in $\mathrm{ba}(\Omega,\mathcal{A})$.

The following theorem says that compactness of $\mathcal{P}$ (with respect to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-topology on $\mathrm{ca}(\Omega,\mathcal{A})$) can be characterized by the relationship between the structure and the credal set of $\overline{P}$:

Theorem 2.27 *Let $\overline{P}$ be an upper expectation on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. Let $\mathcal{P}$ be the structure and $\mathcal{M}$ be the credal set of $\overline{P}$. Then:*
$\mathcal{P}$ is compact (with respect to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-topology on $\mathrm{ca}(\Omega,\mathcal{A})$) if and only if

$$\mathcal{P} \,=\, \mathcal{M} \tag{2.19}$$

Proof: According to Lemma 8.25, the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-topology on $\mathrm{ca}(\Omega,\mathcal{A})$ coincides with the subspace topology on $\mathrm{ca}(\Omega,\mathcal{A})$ generated by the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-topology on $\mathrm{ba}(\Omega,\mathcal{A})$. Hence, $\mathcal{P} \subset \mathrm{ca}(\Omega,\mathcal{A})$ is $\mathcal{L}_\infty(\Omega,\mathcal{A})$-compact in $\mathrm{ca}(\Omega,\mathcal{A})$ if and only if it is $\mathcal{L}_\infty(\Omega,\mathcal{A})$-compact in $\mathrm{ba}(\Omega,\mathcal{A})$. According to Proposition 2.21, $\mathcal{P}$ is $\mathcal{L}_\infty(\Omega,\mathcal{A})$-compact in $\mathrm{ba}(\Omega,\mathcal{A})$ if and only if (2.19); cf. also Corollary 2.16. □

If we explicitly insist on σ-additivity (as in the concept of upper expectations) and consider upper expectations as an independent concept, the $\mathcal{L}_\infty(\Omega,\mathcal{A})$-topology is not a very interesting topology for theoretical investigations. This is a consequence of Theorem 2.27 because it says: A structure is compact only in these cases where there is absolutely no difference between the two concepts "upper expectations" and "coherent upper previsions".

However, cases where structures are compact (with respect to this topology) are very important for applications; cf. Augustin (1998). Of course, Theorem 2.27 does not provide us with a practical criterion in order to check compactness for real applications. In case of F-probabilities, such (necessary and sufficient) criteria are given by (Augustin, 1998, Proposition 2.11). The following theorem is a slight generalization of (Augustin, 1998, Proposition 2.11) – it is not only formulated for F-probabilities but also for general upper expectations; the proof is similar.

Theorem 2.28 *Let $\mathcal{P}$ be the structure of an upper expectation $\overline{P}$ on $\mathcal{L}_\infty(\Omega,\mathcal{A})$. Then the following conditions are all equivalent:*

a) *$\overline{P}$ is continuous.*

b) *$\mathcal{P}$ is uniformly dominated.*

c) *$\mathcal{P}$ is compact with respect to the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology on* $\mathrm{ca}(\Omega,\mathcal{A})$

Proof: Let $\mathcal{M}$ be the credal set of $\overline{P}$; cf. Proposition 2.23.

[**a)**⇒**c)**] Take any $P \in \mathcal{M}$. Then, for every sequence $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ such that $(A_n)_{n\in\mathbb{N}} \searrow \emptyset$, continuity of $\overline{P}$ implies

$$0 \leq \lim_{n\to\infty} P(A_n) \leq \lim_{n\to\infty} \overline{P}\big[I_{A_n}\big] = 0$$

Hence, $P \in \mathrm{ca}(\mathcal{A}) \cap \mathcal{M} = \mathcal{P}$. That is, $\mathcal{M} = \mathcal{P}$ and c) follows from Theorem 2.27.

[**c)**⇒**b)**] Confer (Baumann, 1968, Korollar 2.5).

[**b)**⇒**a)**] is a direct consequence of the definitions.

□

While the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology is quite common on $\mathrm{ba}(\Omega,\mathcal{A})$, it is not very usual on $\mathrm{ca}(\Omega,\mathcal{A})$ and on the set of all probability measures $\mathrm{ca}_1^+(\Omega,\mathcal{A})$. This is underlined by Theorem 2.27 in the context of upper previsions because it says that the $\mathcal{L}_\infty(\Omega,\mathcal{A})$ - topology is most appropriate on $\mathrm{ca}(\Omega,\mathcal{A})$ in these cases where it does not matter if we consider $\mathrm{ca}(\Omega,\mathcal{A})$ or $\mathrm{ba}(\Omega,\mathcal{A})$.
The most common weak topologies on $\mathrm{ca}(\Omega,\mathcal{A})$ are Γ-topologies (cf. Subsection 8.2) where Γ is a certain class of continuous functions on Ω and $\Omega = \Xi$ is a Polish space or a compact Hausdorff space. These topologies are studied in connection with upper expectations in the following two subsections.

2.4.3 Upper expectations on Polish spaces

Let Ξ be a Polish space. That is, Ξ is a topological space and there is a metric d on Ξ such that

- d generates the topology on Ξ
- Ξ is complete and separable with respect to d

Let $\mathfrak{B}$ be the Borel-σ-algebra on Ξ. That is, $\mathfrak{B}$ is the smallest σ-algebra which contains all open sets $V \subset \Xi$.
Let $\mathcal{C}_{\mathrm{b}}(\Xi)$ be the set of all bounded, continuous functions

$$f:\ \Xi \ \to\ \mathbb{R}$$

Especially, $\mathcal{C}_{\mathrm{b}}(\Xi)$ is a norm-closed vector subspace of $\mathcal{L}_\infty(\Xi,\mathfrak{B})$.
Hence, every bounded, signed measure $\mu \in \mathrm{ca}(\Xi,\mathfrak{B})$ uniquely defines a (norm-)continuous linear functional

$$\mu:\ \mathcal{C}_{\mathrm{b}}(\Xi)\ \to\ \mathbb{R}\,, \qquad f\ \mapsto\ \int f\,d\mu\ =\ \mu[f]$$

Note[7] that two different bounded signed measures do not define the same functional on $\mathcal{C}_{\mathrm{b}}(\Xi)$:

$$\mu_1,\, \mu_2 \;\in\; \mathrm{ca}(\Xi,\mathfrak{B})\,, \quad \mu_1[f] \;=\; \mu_2[f] \;\;\forall\, f\in\mathcal{C}_{\mathrm{b}}(\Xi) \qquad\Rightarrow\qquad \mu_1=\mu_2 \tag{2.20}$$

As a consequence, $\mathrm{ca}(\Xi,\mathfrak{B})$ can be identified with a linear subspace in the dual space of $\mathcal{C}_{\mathrm{b}}(\Xi)$.

Analogously to Section 2.3, $\mathrm{ca}(\Xi,\mathfrak{B})$ can be endowed with the $\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology; cf. Subsection 8.2. This is the smallest topology on $\mathrm{ca}(\Xi,\mathfrak{B})$ such that

$$\mathrm{ca}(\Xi,\mathfrak{B}) \;\to\; \mathbb{R}\,, \qquad \mu \;\mapsto\; \mu[f]$$

is continuous for every $f\in\mathcal{C}_{\mathrm{b}}(\Xi)$.

However, it is more common to use the "$\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology on $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$" [8] in classical probability theory. This is the smallest topology on the set of all probability measures $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ such that

$$\mathrm{ca}_1^+(\Xi,\mathfrak{B}) \;\to\; \mathbb{R}\,, \qquad \mu \;\mapsto\; \mu[f]$$

is continuous for every $f\in\mathcal{C}_{\mathrm{b}}(\Xi)$. This topology is called *weak topology of probability measures.* A standard reference for this topology is Billingsley (1968). The use of the weak topology of probability measures is very pleasant because it is metrizable so that it suffices to consider sequences instead of nets.

Furthermore, there is a simple characterization of relatively compact sets:

Theorem 2.29 (Prohorov) *A subset of* $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ *is relatively compact in the weak topology of probability measures if and only if it is tight.*

Confer (Billingsley, 1968, Theorem 6.1).

Finally, there is, of course, also a strong connection between this topology and the $\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology on $\mathrm{ca}(\Xi,\mathfrak{B})$. The following lemma implies that a subset of $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ is closed in the weak topology of probability measures if and only if it is closed in the $\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology on $\mathrm{ca}(\Xi,\mathfrak{B})$. This is needed in the proof of Proposition 2.33 which characterizes weakly closed subsets of $\mathrm{ca}(\Xi,\mathfrak{B})$.

Lemma 2.30 $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ *is* $\mathcal{C}_{\mathrm{b}}(\Xi)$ - *closed in* $\mathrm{ca}(\Xi,\mathfrak{B})$.

Proof: Let $(P_\beta)_{\beta\in B}$ be a net in $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ such that

$$P_\beta \;\xrightarrow{\beta}\; \mu \;\in\; \mathrm{ca}(\Xi,\mathfrak{B}) \qquad \text{in the } \mathcal{C}_{\mathrm{b}}(\Xi)\text{ - topology}$$

Hence, $\mu(\Xi) \;=\; 1$ and $\mu[f] \;\geq\; 0 \;\;\forall\, f\in\mathcal{C}_{\mathrm{b}}(\Xi)$. That is, $\mu \;\in\; \mathrm{ca}_1^+(\Xi,\mathfrak{B})$. □

In the following, the weak topology of probability measures is simply called *weak topology* and topological notions such as "weakly closed" and "converges weakly" are with respect to this topology.

Upper expectations have originally been defined in Buja (1984). There, they are only considered on Polish spaces where the weak topology is used. This setup seems to be promising because of the following statement which is implicitly contained in (Buja, 1984, Proposition 2.1 and 2.2):

[7] It follows from $\mu_1[f]=\mu_2[f]\;\;\forall\, f\in\mathcal{C}_{\mathrm{b}}(\Xi)$ that $\mu_1^+[f]+\mu_2^-[f] \;=\; \mu_2^+[f]+\mu_1^-[f]\;\;\forall\, f\in\mathcal{C}_{\mathrm{b}}(\Xi)$. Hence, $\mu_1^+ + \mu_2^- \;=\; \mu_2^+ + \mu_1^-$ according to (Bauer, 2001, Lemma 30.14) and, therefore, $\mu_1=\mu_2$.

[8] This topology is not defined in Subsection 8.2 because $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ is not a linear space.

"Every tight structure is weakly compact."

However, this is, in general, not correct! For a counterexample, confer Subsection 4.1.2. In the following, two theorems are given which provide necessary and sufficient conditions for relative weak compactness and weak compactness respectively. It is surprising that one of these conditions – namely condition a) in both theorems – is very similar to continuity which is a necessary and sufficient condition for $\mathcal{L}_\infty(\Xi,\mathfrak{B})$ - compactness in $\mathrm{ca}(\Xi,\mathfrak{B})$; cf. Theorem 2.28. [9]

Theorem 2.31 *Let $\mathcal{P}$ be the structure of an upper expectation $\overline{P}$. Then, the following conditions are all equivalent:*

a) *If $(f_n)_{n\in\mathbb{N}}\subset\mathcal{C}_\mathrm{b}(\Xi)$ is a sequence such that $f_n(x)\searrow f(x)\quad\forall\, x\in\Xi$ for some $f\in\mathcal{C}_\mathrm{b}(\Xi)$, then $\overline{P}[f_n]\searrow\overline{P}[f]$.*

b) *$\mathcal{P}$ is tight.*

c) *$\mathcal{P}$ is relatively weakly compact in $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$.*

Proof:

[**a)** $\Rightarrow$ **b)**] Let d be a metric on Ξ so that d induces the topology of Ξ and (Ξ,d) becomes a complete separable metric space. Let $(a_n)_{n\in\mathbb{N}}$ be a dense subset of Ξ. For every $r>0$ define

$$\gamma_i^r\ :\ \Xi\longrightarrow\mathbb{R}\,,\quad x\mapsto\frac{d(x,a_i)\wedge r}{r}\,,\qquad i\in\mathbb{N}$$

$$\Gamma_n^r:=\min_{i=1,\dots,n}\gamma_i^r\,,\qquad n\in\mathbb{N}$$

$$\overline{B}_r(a_i)=\big\{x\in\Xi\ \big|\ d(x,a_i)\le r\big\}$$

and $\overline{B}_r^{\,\mathrm{c}}(a_i)=\Xi\setminus\overline{B}_r(a_i)$. Then:

[1] $\Gamma_n^r\ge I_{\bigcap_{i=1}^n\overline{B}_r^{\,\mathrm{c}}(a_i)}\quad\forall\, n\in\mathbb{N},\ \forall\, r>0$:

Since $\gamma_i^r\ge I_{\overline{B}_r^{\,\mathrm{c}}(a_i)}\ \ \forall\, i\in\mathbb{N}$,

$$\Gamma_n^r\ =\ \min_{i=1,\dots,n}\gamma_i^r\ \ge\ \min_{i=1,\dots,n}I_{\overline{B}_r^{\,\mathrm{c}}(a_i)}\ =\ I_{\bigcap_{i=1}^n\overline{B}_r^{\,\mathrm{c}}(a_i)}$$

[2] $\forall\,\varepsilon>0,\ \ \forall\, k\in\mathbb{N}\quad\exists\, n_k\in\mathbb{N}:\ 0\ \le\ \overline{P}\big[\Gamma_{n_k}^{1/k}\big]\ \le\ \varepsilon\cdot 2^{-k}$:

Since $(a_n)_{n\in\mathbb{N}}$ is d-dense in Ξ, $\Gamma_n^{1/k}\searrow 0$ pointwise for $n\to\infty$. Then it follows from a), that

$$\lim_n\overline{P}\Big[\Gamma_n^{1/k}\Big]\searrow\overline{P}[0]=0$$

because $(\Gamma_n^{1/k})_{n\in\mathbb{N}}\subset\mathcal{C}^b(\Xi,\mathfrak{B})$.

[9] In case of locally compact separable metric spaces (i.e. in a special case of Polish spaces), it has also been discovered in Lasserre (1998) that such "continuity conditions" serve as a uniform principle in order to characterize (sequentially) compact sets of probability measures with respect to several topologies.

[3] $K_\varepsilon = \bigcap_{k\in\mathbb{N}}\bigcup_{i=1}^{n_k}\overline{B}_{1/k}(a_i)$ is obviously totally bounded. Furthermore, K_ε is closed and complete. Hence, K_ε is compact (cf. (Dunford and Schwartz, 1958, Theorem I.6.15)) and we have

$$\begin{aligned}\overline{P}\big(\Xi\setminus K_\varepsilon\big) \;&=\; \overline{P}\Big(\bigcup_{k\in\mathbb{N}}\bigcap_{i=1}^{n_k}\overline{B}^{\,\mathrm{c}}_{1/k}(a_i)\Big) \;\le\; \sum_k \overline{P}\Big(\bigcap_{i=1}^{n_k}\overline{B}^{\,\mathrm{c}}_{1/k}(a_i)\Big) \;\le\\ &\overset{[1]}{\le}\; \sum_k \overline{P}\big[\Gamma^{1/k}_{n_k}\big] \;\overset{[2]}{\le}\; \sum_k \varepsilon\cdot 2^{-k} \;=\; \varepsilon\end{aligned}$$

[b) ⇒ a)] Let $(f_n)_{n\in\mathbb{N}}\subset\mathcal{C}_{\mathrm{b}}(\Xi)$ be a sequence such that $f_n(x)\searrow f(x)\ \ \forall x\in\Xi$ for some $f\in\mathcal{C}_{\mathrm{b}}(\Xi)$. For each $l\in\mathbb{N}$, we may choose a compact $K_l\subset\Xi$, so that

$$\overline{P}\big(\Xi\setminus K_l\big) < \frac{1}{l} \tag{2.21}$$

because $\mathcal{P}$ is assumed to be tight. According to Dini's Theorem (Dudley, 1989, Theorem 2.4.10), $(f_n)_{n\in\mathbb{N}}$ converges uniformly on every compact set K_l. Hence, for every $l\in\mathbb{N}$,

$$\begin{aligned}&\limsup_n \big|\overline{P}[f_n]-\overline{P}[f]\big| \;\le\; \limsup_n \sup_{P\in\mathcal{P}} P[f_n-f] \;\le\\ &\le\; \limsup_n \Big(\sup_{P\in\mathcal{P}}\int_{K_l} f_n - f\,dP + \sup_{P\in\mathcal{P}}\int_{\Xi\setminus K_l} f_n - f\,dP\Big) \;\le\\ &\le\; \limsup_n \Big(\sup_{x\in K_l}(f_n-f) + \big(\|f_n\|_\infty+\|f\|_\infty\big)\overline{P}(\Xi\setminus K_l)\Big) \;\le\\ &\le\; 2\|f_1\|_\infty\cdot\frac{1}{l}\end{aligned}$$

[c) ⇔ b)] This is the content of Theorem 2.29.

□

Theorem 2.31 cannot be proven by a simple application of (Föllmer and Schied, 2002, Theorem 3.8) since (Föllmer and Schied, 2002, Theorem 3.8) is not correct as explained in (Föllmer and Schied, 2004, Remark 4.29). Nevertheless, the proof of Theorem 2.31 is only a variant of the proof of (Föllmer and Schied, 2002, Theorem 3.8). Theorem 2.31 essentially coincides also with (Varadarajan, 1965, Theorem II.25).

Theorem 2.32 *Let $\mathcal{P}$ be the structure of an upper expectation $\overline{P}$. Then the following conditions are all equivalent:*

a) *If $(f_n)_{n\in\mathbb{N}}\subset\mathcal{C}_{\mathrm{b}}(\Xi)$ is a sequence such that $f_n(x)\searrow f(x)\ \ \forall x\in\Xi$ for some $f\in\mathcal{L}_\infty(\Xi,\mathfrak{B})$, then $\overline{P}[f_n]\searrow\overline{P}[f]$.*

b) *$\mathcal{P}$ is tight and weakly closed in $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$.*

c) *$\mathcal{P}$ is weakly compact in $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$.*

Proof:

[a)⇒b)]

[1] Tightness of $\mathcal{P}$ follows from Theorem 2.31.

[2] $\mathcal{P}$ is weakly closed in $\mathrm{ca}_1^+(\Xi, \mathfrak{B})$:
Let $(P_k)_{k\in\mathbb{N}} \subset \mathcal{P}$ be any sequence which converges weakly to some $P \in \mathrm{ca}_1^+(\Xi, \mathfrak{B})$. Since the weak topology on $\mathrm{ca}_1^+(\Xi, \mathfrak{B})$ is metrizable, it suffices to show that $P \in \mathcal{P}$. To this end, let f be any upper semi-continuous function in $\mathcal{L}_\infty(\Xi, \mathfrak{B})$. Then, there is a sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}_\mathrm{b}(\Xi)$ such that $f_n(x) \searrow f(x)\ \ \forall x \in \Xi$ (cf. (Denkowski et al., 2003, Proposition 1.4.54)). For every $g \in \mathcal{C}_\mathrm{b}(\Xi)$, $P[g] = \lim_k P_k[g] \leq \overline{P}[g]$. Hence, the dominated convergence theorem implies

$$P[f] = \lim_n P[f_n] \leq \lim_n \overline{P}[f_n] \overset{a)}{=} \overline{P}[f]$$

That is, $P[f] \leq \overline{P}[f]$ for every upper semi-continuous function in $\mathcal{L}_\infty(\Xi, \mathfrak{B})$.
Now, let f be any function in $\mathcal{L}_\infty(\Xi, \mathfrak{B})$. According to Lemma 8.27, there is a sequence $(f_n)_{n\in\mathbb{N}}$ of upper semi-continuous function $f_n \in \mathcal{L}_\infty(\Xi, \mathfrak{B})$ such that

$$f_1 \leq f_2 \leq f_3 \leq \ldots \leq f \qquad \text{and} \qquad P[f_n] \nearrow P[f] \tag{2.22}$$

Hence,

$$P[f] \overset{(2.22)}{=} \lim_{n\to\infty} P[f_n] \leq \overline{P}[f_n] \overset{(2.22)}{\leq} \overline{P}[f]$$

[b)⇒c)] follows from Theorem 2.29.

[c)⇒a)] Let $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}_\mathrm{b}(\Xi)$ be a sequence such that $f_n(x) \searrow f(x)\ \ \forall x \in \Xi$ for some $f \in \mathcal{L}_\infty(\Xi, \mathfrak{B})$. Then:

[1] $\lim_n \overline{P}[f_n]$ exists because $\overline{P}[f_1] \geq \overline{P}[f_2] \geq \cdots \geq \overline{P}[f] > -\infty$

[2] For every $n \in \mathbb{N}$, there is some $P_n \in \mathcal{P}$ so that $P_n[f_n] = \overline{P}[f_n]$:
The weakly continuous functions

$$\mathcal{P} \to \mathbb{R}\,, \qquad P \mapsto P[f_n]$$

attain their maxima $\overline{P}[f_n]$ in some $P_n \in \mathcal{P}$ because $\mathcal{P}$ is weakly compact (cf. (Denkowski et al., 2003, Theorem 1.3.11)).

[3] There is a subsequence $(P_{n_l})_{l\in\mathbb{N}}$ which converges weakly to some $P_0 \in \mathcal{P}$ because $\mathcal{P}$ is also sequentially weakly compact (Metrizability!).

[4] $\lim_l P_{n_l}[f_{n_l}] \leq P_0[f]$:
The limit $\lim_l P_{n_l}[f_{n_l}]$ exists according to the definition of $(P_n)_{n\in\mathbb{N}}$ and part [1] of the present proof. The sequence $(f_n)_{n\in\mathbb{N}}$ is decreasing. Hence for every $k \in \mathbb{N}$, $\lim_l P_{n_l}[f_{n_l}] \leq \lim_l P_{n_l}[f_k] \overset{[3]}{=} P_0[f_k]$ and the dominated convergence theorem implies

$$\lim_l P_{n_l}[f_{n_l}] \leq \lim_k P_0[f_k] = P_0[f]$$

[5] Finally, $\overline{P}[f] \overset{[1]}{\leq} \lim_n \overline{P}[f_n] = \lim_l \overline{P}[f_{n_l}] \overset{[2]}{=} \lim_l P_{n_l}[f_{n_l}] \overset{[4]}{\leq} P_0[f] \overset{[3]}{\leq} \overline{P}[f]$

□

Assumption a) of Theorem 2.31 and assumption a) of Theorem 2.32 seem to be nearly the same. But the difference matters as can be seen from Theorem 2.31, Theorem 2.32 and Subsection 4.1.2.

In addition, the next proposition characterizes weak closedness of structures.

Proposition 2.33 *The structure $\mathcal{P}$ of an upper expectation $\overline{P}$ is weakly closed if and only if it has the following representation:*

$$\mathcal{P} \;=\; \big\{ P \in \mathrm{ca}_1^+(\Xi, \mathfrak{B}) \;\big|\; P[f] \leq \overline{P}[f] \;\; \forall f \in \mathcal{C}_{\mathrm{b}}(\Xi) \big\} \tag{2.23}$$

Proof: Put $\mathcal{V} = \mathcal{P}$, $\Gamma = \mathcal{C}_{\mathrm{b}}(\Xi)$ and $M = \mathrm{ca}(\Xi, \mathfrak{B})$. Since $\mathcal{P}$ is convex, it follows from Theorem 8.26 that the structure $\mathcal{P}$ is $\mathcal{C}_{\mathrm{b}}(\Xi)$ - closed in $\mathrm{ca}(\Xi, \mathfrak{B})$ if and only if it has the representation (2.23).

Hence, it only remains to show that $\mathcal{P} \subset \mathrm{ca}_1^+(\Xi, \mathfrak{B})$ is $\mathcal{C}_{\mathrm{b}}(\Xi)$ - closed in $\mathrm{ca}(\Xi, \mathfrak{B})$ if and only if it is weakly closed in $\mathrm{ca}_1^+(\Xi, \mathfrak{B})$.
Let $\mathcal{P}$ be $\mathcal{C}_{\mathrm{b}}(\Xi)$ - closed in $\mathrm{ca}(\Xi, \mathfrak{B})$ and $(P_\beta)_{\beta \in B} \subset \mathcal{P}$ be a net such that

$$P_\beta \xrightarrow{\beta} P \;\in\; \mathrm{ca}_1^+(\Xi, \mathfrak{B}) \qquad \text{weakly}$$

That is, $P_\beta[f] \;\longrightarrow\; P[f]$ for every $f \in \mathcal{C}_{\mathrm{b}}(\Xi)$ and, therefore, $P_\beta \;\longrightarrow\; P$ in the $\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology. Finally, it follows from $\mathcal{C}_{\mathrm{b}}(\Xi)$ - closedness of $\mathcal{P}$ that $P \in \mathcal{P}$.
Conversely, let $\mathcal{P}$ be weakly closed in $\mathrm{ca}_1^+(\Xi, \mathfrak{B})$ and $(P_\beta)_{\beta \in B} \subset \mathcal{P}$ be a net such that

$$P_\beta \xrightarrow{\beta} \mu \;\in\; \mathrm{ca}(\Xi, \mathfrak{B}) \qquad \text{in the } \mathcal{C}_{\mathrm{b}}(\Xi)\text{ - topology}$$

According to Lemma 2.30, $\mu \in \mathrm{ca}_1^+(\Xi, \mathfrak{B})$. Therefore, convergence in the $\mathcal{C}_{\mathrm{b}}(\Xi)$ - topology implies weak convergence. Finally, it follows from weak closedness of $\mathcal{P}$ that $\mu \in \mathcal{P}$. □

Proposition 2.33 seems to indicate that this setup (weak topology and Polish spaces) is not appropriate because it says that an upper expectation $\overline{P}$ whose structure is $\mathcal{C}_{\mathrm{b}}(\Xi)$ - closed (or compact) is at least implicitly defined by its values on $\mathcal{K} = \mathcal{C}_{\mathrm{b}}(\Xi)$. However, one of the most important special cases of upper expectations are F-probabilities which are defined by their values on some $\mathcal{K} \subset \big\{ I_B \;\big|\; B \in \mathfrak{B} \big\}$ and, at least in case of $\Xi = \mathbb{R}$, indicator functions are rarely continuous.

The following subsection investigates upper expectations on compact Hausdorff spaces – a setup which seems to be similar to the Polish setup. However, compact Hausdorff spaces prove to be much more important (at least from a theoretical point of view) as will be seen in Subsection 2.5.

2.4.4 Upper expectations on compact Hausdorff spaces

In this subsection, we turn over to compact Hausdorff spaces. While upper expectations on Polish spaces have been considered in Buja (1984), upper expectations have not been studied explicitly on compact Hausdorff spaces before. In Subsection 2.5, this setup turns out to be important and – in a sense – as general as coherent upper previsions on arbitrary spaces.
Let Ξ be a compact Hausdorff space – i.e. Ξ is a topological Hausdorff space which is compact. Let $\mathcal{C}(\Xi)$ be the set of all continuous functions

$$f : \;\; \Xi \;\longrightarrow\; \mathbb{R}$$

and let $\mathfrak{B}_0$ be the Baire-σ-algebra on Ξ. This is the smallest σ-algebra on Ξ such that every continuous function $f \in \mathcal{C}(\Xi)$ is measurable. Obviously, the Baire-σ-algebra is

contained in the Borel-σ-algebra: $\mathfrak{B}_0 \subset \mathfrak{B}$. In any metric space, the two σ-algebras coincide $\mathfrak{B}_0 = \mathfrak{B}$ (for the metric topology); cf. (Dudley, 1989, Theorem 7.1.1).
Compactness of Ξ implies that every continuous function $f \in \mathcal{C}(\Xi)$ is bounded. So, $\mathcal{C}(\Xi) = \mathcal{C}_\mathrm{b}(\Xi)$ and $\mathcal{C}(\Xi)$ is a norm-closed vector subspace of $\mathcal{L}_\infty(\Xi, \mathfrak{B}_0)$.
Every bounded, signed measure $\mu \in \mathrm{ca}(\Xi, \mathfrak{B}_0)$ uniquely defines a (norm-)continuous linear functional

$$T_\mu : \ \mathcal{C}_\mathrm{b}(\Xi) \ \to \ \mathbb{R}\,, \qquad f \ \mapsto \ \int f\,d\mu \ = \ \mu[f]$$

The converse statement is also true; cf. (Dudley, 1989, Theorem 7.4.1):

Theorem 2.34 *For every $\mu \in \mathrm{ca}(\Xi, \mathfrak{B}_0)$,*

$$T_\mu : \ \mathcal{C}(\Xi) \ \to \ \mathbb{R}\,, \qquad f \ \mapsto \ \int f\,d\mu$$

is a continuous linear functionals on $\mathcal{C}(\Xi)$.
Conversely, for every continuous linear functional

$$T : \ \mathcal{C}(\Xi) \ \to \ \mathbb{R}\,, \qquad f \ \mapsto \ T(f)$$

on $\mathcal{C}(\Xi)$, there is a unique $\mu \in \mathrm{ca}(\Omega, \mathcal{A})$ such that

$$T(f) \ = \ \int f\,d\mu \qquad f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$$

Remark 2.35 *$\mu \ \mapsto \ T_\mu$ is a vector space isomorphism between $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ and the dual space of $\mathcal{C}(\Xi)$ denoted by $\mathcal{C}(\Xi)^*$. Furthermore, $\mu \ \mapsto \ T_\mu$ is isometric: $\|T_\mu\|_* \ = \ \|\mu\|$ where $\|\cdot\|_*$ denotes the dual norm $\|T\|_* \ = \ \sup\big\{T(f)\ \big|\ \|f\| \le 1,\ f \in \mathcal{C}(\Xi)\big\}$.*
It can also easily be read off from the proof of Theorem 2.34 in (Dudley, 1989, Theorem 7.4.1) that this isomorphism respects order:

$$\mu \ \ge \ 0 \qquad \Leftrightarrow \qquad T_\mu(f) \ \ge \ 0 \quad \forall f \ge 0\,, \ \ f \in \mathcal{C}(\Xi) \tag{2.24}$$

Summing up, $\mu \ \mapsto \ T_\mu$ is an L-space isomorphism between $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ and $\mathcal{C}(\Xi)^*$.

Since $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ can be identified with the dual space of $\mathcal{C}(\Xi)$, $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ is usually provided with the Γ-topology (cf. Subsection 8.2) where $\Gamma = \mathcal{C}(\Xi)$. [10] This is the weakest topology on $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ such that, for every $f \in \mathcal{C}(\Xi)$,

$$\Lambda_f \ : \ \ \mathrm{ca}(\Xi, \mathfrak{B}_0) \ \to \ \mathbb{R}\,, \qquad \mu \ \mapsto \ \Lambda_f(\mu) \ = \ \mu[f]$$

is continuous.

The following theorem provides necessary and sufficient conditions for $\mathcal{C}(\Xi)$ - compactness of structures.

Theorem 2.36 *Let $\mathcal{P}$ be the structure of an upper expectation $\overline{P}$ on $(\Xi, \mathfrak{B}_0)$. Then, the following statements are all equivalent:*

a) *$\mathcal{P}$ is $\mathcal{C}(\Xi)$ - compact.*

[10] In the more general case where Ξ is a locally compact Hausdorff space, the usual weak topology would be given by $\Gamma = \mathcal{C}_0(\Xi)$, the set of all continuous functions $f \in \mathcal{C}(\Xi)$ with compact support. However, $\mathcal{C}(\Xi) = \mathcal{C}_0(\Xi)$ in case of a compact Hausdorff space Ξ.

b) *$\mathcal{P}$ is $\mathcal{C}(\Xi)$ - closed.*

c) *$\mathcal{P}$ can be written as*

$$\mathcal{P} \;=\; \big\{P \in \mathrm{ca}_1^+(\Xi, \mathfrak{B}_0) \;\big|\; P[f] \le \overline{P}[f] \quad \forall\, f \in \mathcal{C}(\Xi)\big\} \tag{2.25}$$

Proof:

$(b) \Leftrightarrow (c)$: Since $\mathcal{P}$ is convex, this follows from Theorem 8.26 where $\mathcal{V} = \mathcal{P}$, $\Gamma = \mathcal{C}(\Xi)$ and $M = \mathrm{ca}(\Xi, \mathfrak{B}_0)$.

$(a) \Rightarrow (b)$: Ξ is assumed to be a Hausdorff space.

$(a) \Leftarrow (b)$: According to Theorem 2.34 and Remark 2.35, $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ can be identified with the dual space of $\mathcal{C}(\Xi)$. Hence, the closed unit sphere $\{\mu \in \mathrm{ca}(\Xi, \mathfrak{B}_0) \mid \|\mu\| \le 1\}$ is $\mathcal{C}(\Xi)$ - compact in $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ according to (Dunford and Schwartz, 1958, Theorem V.4.2). Since $\mathcal{P}$ is a $\mathcal{C}(\Xi)$ - closed subset of this $\mathcal{C}(\Xi)$ - compact set, $\mathcal{P}$ is $\mathcal{C}(\Xi)$ - compact, too. □

Analogously to Polish spaces, one could argue that Theorem 2.36 indicates that this setup ($\mathcal{C}(\Xi)$ - topology and compact spaces) is not appropriate because it says that an upper expectation $\overline{P}$ whose structure is $\mathcal{C}(\Xi)$ - compact is at least implicitly defined by its values on $\mathcal{K} = \mathcal{C}(\Xi)$. This is true from a practically orientated point of view – however, it is not true from a theoretical point of view as can be seen in Section 2.5.

2.4.5 F-probabilities

As stated before, F-probabilities are the most important and also most investigated special case of upper expectations. In Subsection 2.4.1, F-probabilities have already been defined as special cases of upper expectations. Since the definition of F-probabilities is usually not given in terms of upper expectations and there is a different notation for F-probabilities, the usual definitions and notations of F-Probabilities are given in the present subsection so that the connection to upper expectations gets more visible.

The following definitions originate from Weichselberger (2000). Confer also Weichselberger (2001).

Definition 2.37 *Let $(\Omega, \mathcal{A})$ be a measurable space.*

a) *A function p on $\mathcal{A}$ satisfying the axioms of Kolmogorov is called* classical probability*. That is, the set of all classical probabilities on $(\Omega, \mathcal{A})$ is $\mathrm{ca}_1^+(\Omega, \mathcal{A})$.*

b) *A function P of the form*

$$P : \mathcal{A} \to \{[L, U] \mid \; 0 \le L \le U \le 1\}, \qquad A \mapsto P(A) = [L(A), U(A)]$$

is called R-probability *with* structure $\mathcal{M}$ *if the set*

$$\mathcal{M} := \{p \in \mathrm{ca}_1^+(\Omega, \mathcal{A}) \mid \; L(A) \le p(A) \le U(A) \;\; \forall\, A \in \mathcal{A}\}$$

is not empty. L is called lower probability *and U is called* upper probability*.*

c) *An* R*-probability*

$$P : \mathcal{A} \to \{[L, U] \mid \; 0 \le L \le U \le 1\}, \qquad A \mapsto P(A) = [L(A), U(A)]$$

is called F-probability *if*

$$\left.\begin{array}{r} \inf_{p \in \mathcal{M}} p(A) = L(A) \\ \sup_{p \in \mathcal{M}} p(A) = U(A) \end{array}\right\} \quad \forall\, A \in \mathcal{A}$$

For every F-probability, L and U are conjugate. i.e.

$$L(A) = 1 - U(A^c) \qquad \forall\, A \in \mathcal{A}$$

Therefore, every F-probability is uniquely determined by $L : \mathcal{A} \to [0,1]$. $(\Omega, \mathcal{A}, L)$ is called *F-probability field.*

The following definition extends the concept of expectations in case of classical probability to interval probability:

Definition 2.38 *For every F-probability field* $(\Omega, \mathcal{A}, L)$ *with structure* $\mathcal{M}$*, a random variable* $X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathbb{B})$ *is called* $\mathcal{M}$-integrable *if* X *is p-integrable for each element p of* $\mathcal{M}$*. Then*

$$\mathbb{E}_{\mathcal{M}}X := \big[L\mathbb{E}_{\mathcal{M}}X, U\mathbb{E}_{\mathcal{M}}X\big] := \left[\inf_{p \in \mathcal{M}} \mathbb{E}_p X, \sup_{p \in \mathcal{M}} \mathbb{E}_p X\right] \subset [-\infty, \infty]$$

is called the (interval-valued) expectation of X *(with respect to* $(\Omega, \mathcal{A}, L)$*).*

Let $(\Omega, \mathcal{A}, L)$ be an F-probability field with structure $\mathcal{M}$. Put

$$\overline{P}[I_A] \;:=\; U(A) \qquad \forall\, A \in \mathcal{A}$$

Obviously,

$$\overline{P} \;:\; \big\{I_A \,\big|\; A \in \mathcal{A}\big\} \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; \overline{P}[I_A]$$

is an upper expectation with structure $\mathcal{M}$. Furthermore, $\overline{P}$ is indeed an F-probability according to Definition 2.18 where

$$\mathcal{K} \;=\; \big\{I_A \,\big|\; A \in \mathcal{A}\big\} \tag{2.26}$$

Definition 2.18 also admits

$$\mathcal{K} \;\subsetneq\; \big\{I_A \,\big|\; A \in \mathcal{A}\big\} \tag{2.27}$$

instead of (2.26). In this case, F-probabilities are called *partially determinate F-probabilities* in Weichselberger (2000) and Weichselberger (2001).
According to Proposition 2.19, an F-probability $\overline{P}$ can be extended to an upper expectation on $\mathcal{L}_\infty(\Omega, \mathcal{A})$. This extension is equal to

$$\overline{P}[f] \;=\; U\mathbb{E}_{\mathcal{M}}f \quad \forall\, f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$$

(Of course, every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$ is $\mathcal{M}$-integrable.)

Caution 1: *Here, the structure is denoted by* $\mathcal{M}$ *in order to be in line with the notation used in Weichselberger (2001). In general, this* $\mathcal{M}$ *is* <u>*not*</u> *the credal set of* $\overline{P}$ *according to Proposition 2.21. In the previous subsections, structures are denoted by* $\mathcal{P}$ *because* $\mathcal{M}$ *already denotes credal sets in the theory of imprecise probabilities according to Walley (1991) (which is mainly used in the present book).*
Caution 2: *As stated before, the extension of a (partially determinate) F-probability to an upper expectation on* $\mathcal{L}_\infty(\Omega, \mathcal{A})$ *is equal to*

$$\overline{P}[f] \;=\; U\mathbb{E}_{\mathcal{M}}f \quad \forall\, f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$$

so that we are in accordance with Weichselberger (2000) and Weichselberger (2001). However, one has to be careful when a structure is generated from a prestructure. Then, it makes a difference if the structure of an F-probability or the structure of an upper expectation on $\mathcal{L}_\infty(\Omega, \mathcal{A})$ is generated. Confer page 23.

It is a basic property of classical probability measures that they are uniquely determined by their values on a $\cap$-stable generator of the σ-algebra $\mathcal{A}$. This is not true for F-probabilities. However, there is at least a related result on Polish spaces $(\Xi, \mathfrak{B})$:

Proposition 2.39 *Let Ξ be a Polish space with Borel-σ-algebra $\mathfrak{B}$ and let U be the upper probability in an F-probability field $(\Xi, \mathfrak{B}, L)$ with structure $\mathcal{M}$.*
Then

$$U(B) = \sup\big\{U(K) \;\big|\; K \subset B,\ K \textit{ compact}\big\} \qquad \forall\, B \in \mathcal{B}$$

and

$$P \in \mathcal{M} \quad \Leftrightarrow \quad P(K) \le U(K) \ \textit{ for every compact } K \in \mathcal{B}$$

for every $P \in \mathrm{ca}_1^+(\Xi, \mathfrak{B})$. Especially, an F-probability is uniquely determined by the values of U on the compact subsets of Ξ.

Proof: For $B \in \mathcal{B}$, there is a sequence $(P_n)_{n\in\mathbb{N}} \subset \mathcal{M}$ such that $P_n(B) \nearrow U(B)$. Since Ξ is Polish, every probability measure on $(\Xi, \mathcal{B})$ is regular. Hence,

$$\forall\, n \in \mathbb{N} \ \ \exists\, K_n \subset B,\ K_n \text{ compact}: \ \ 0 \le P_n(B) - P_n(K_n) \le \frac{1}{n}$$

and

$$\begin{aligned}
U(B) &\ge \sup\big\{U(K) \;\big|\; K \subset B,\ K \text{ compact}\big\} \ge \\
&\ge \limsup_n \sup\big\{P_n(K) \;\big|\; K \subset B,\ K \text{ compact}\big\} \ge \\
&\ge \limsup_n P_n(K_n) = \limsup_n \big(P_n(B) - (P_n(B) - P_n(K_n))\big) \ge \\
&\ge \liminf_n P_n(B) - \limsup_n \big(P_n(B) - P_n(K_n)\big) = U(B)
\end{aligned}$$

Let P be a probability measure on $(\Xi, \mathcal{B})$ where $P(K) \le U(K)$ for every compact $K \in \mathcal{B}$. So, P is regular and for every $B \in \mathcal{B}$,

$$\begin{aligned}
P(B) &= \sup\big\{P(K) \;\big|\; K \subset B,\ K \text{ compact}\big\} \le \\
&\le \sup\big\{U(K) \;\big|\; K \subset B,\ K \text{ compact}\big\} = U(B)
\end{aligned}$$

□

2.5 Representation of coherent upper previsions

2.5.1 Introduction

According to the previous sections, there are several ways to define imprecise probabilities. The main difference results from different answers to the question whether to insist on σ-additivity or not. Both the concept of coherent upper previsions and the concept of upper expectations can be based on sets of precise probabilities

$$\big\{P \;\big|\; P[f] \le \overline{P}[f] \ \ \forall\, f\,\big\} \tag{2.28}$$

such that $\overline{P}[f]$ is the supremum of $P[f]$ over this set for every f. If precise probability assignments are modeled by σ-additive probability measures P, we naturally end up with the concept of upper expectations. If precise probability assignments may be modeled by any probability charge P, we naturally end up with the concept of coherent upper previsions.
Therefore, the question whether to use coherent upper previsions or upper expectations corresponds to the question whether precise probability assignments may be modeled by probability charges which are not σ-additive. In case of precise probabilities, it is well known that the answer to this question is hardly connected with the real world. In the spirit of the work of L. Le Cam (cf. e.g. (Le Cam, 1986, Chapter 1)), this is because probability charges and probability measures are no observable objects but they are only mathematical constructs which are intended to represent probability assignments from the real world. In order to do such a representation in a mathematical rigorous way, an "appropriate" sample space $(\Omega, \mathcal{A})$ has to be chosen. However, the Stone representation theorem (Dunford and Schwartz, 1958, § I.12) implies that it only depends on the choice of the sample space whether a precise probability assignment leads to a σ-additive probability measure or not: Even if the precise probability assignment leads to a probability charge which is not σ-additive, there is always another appropriate choice of the sample space which would have led to a σ-additive probability measure. That is, whether we are faced with σ-additive probability measures or not depends on the arbitrary choice of the sample space. For a more detailed explanation of these considerations in the spirit of L. Le Cam, confer Section 3.4.

In short, the present section shows that the same reasoning also applies for imprecise probabilities. That is, whether an imprecise probability assignment leads to an upper expectation (which is based on σ-additivity) or not, only depends on the arbitrary choice of the sample space. As in case of precise probabilities, this is also a consequence of the Stone representation theorem.
In order to see this, Subsection 2.5.2 recalls the Stone representation theorem and presents some preparations based on this fundamental theorem. Next, it is shown in Subsection 2.5.3: Even if an imprecise probability assignment does not lead to an upper expectation but to a coherent upper prevision, there is always another appropriate choice of the sample space such that the imprecise probability assignment leads to an upper expectation. In this way, every coherent upper prevision can be represented by an upper expectation. Since this can be done in a canonical way, we are able to define a "canonical Stone representation" for every coherent upper prevision.
This is not only interesting from a theoretical point of view but also serves as an important tool in Subsection 3.3.3 where standard models are defined by use of σ-additivity. This can be done for every coherent upper prevision via the canonical Stone representation. The results of the present section implies that an analogous proceeding is always possible if σ-additivity is needed in the definition of concepts which originally rely on σ-additivity. For example, this is also possible in order to define conditional coherent upper previsions and this offers an expedient alternative to the definitions based on conglomerability given by Walley (1991).

2.5.2 Stone representation

Let Ω be a set with algebra $\mathcal{A}$. The following famous theorem connects the general setup $(\Omega, \mathcal{A})$ with the setup in Subsection 2.4.4 where Ξ is a compact Hausdorff space and $\mathfrak{B}_0$

the Baire-σ-algebra on Ξ.
A topological space Ξ is called *totally disconnected* if its topology has a base which consists of clopen sets. A set is called *clopen* if it is closed and open.[11] It follows from the definitions that the clopen sets $C \subset \Xi$ form an algebra $\mathfrak{C}$ on Ξ.

Theorem 2.40 (Stone representation theorem)
There is a totally disconnected compact Hausdorff space Ξ such that $\mathcal{A}$ is isomorphic to the algebra $\mathfrak{C}$ of all clopen sets $C \subset \Xi$. That is: There is a bijective map $\Phi : \mathcal{A} \to \mathfrak{C}$ such that

$$\Phi(A_1 \cap A_2) \;=\; \Phi(A_1) \cap \Phi(A_2)\,, \qquad \Phi(A_1 \cup A_2) \;=\; \Phi(A_1) \cup \Phi(A_2)\,,$$
$$\Phi(A^{\mathrm{C}}) \;=\; \big(\Phi(A)\big)^{\mathrm{C}}$$

for every A_1, A_2, $A \in \mathcal{A}$.

For the proof of this theorem, confer e.g. (Dunford and Schwartz, 1958, § I.12).

Note that the properties of the isomorphism $\Phi : \mathcal{A} \to \mathfrak{C}$ imply

$$\Phi(\emptyset) \;=\; \emptyset\,, \qquad \Phi^{-1}(\emptyset) \;=\; \emptyset\,, \qquad \Phi(\Omega) \;=\; \Xi\,, \qquad \Phi^{-1}(\Xi) \;=\; \Omega \tag{2.29}$$

where $\Phi^{-1} : \mathfrak{C} \to \mathcal{A}$ denotes the inverse of Φ.

Φ induces a map $\xi : \mathcal{L}_\infty(\Omega, \mathcal{A}) \to \mathcal{L}_\infty(\Xi, \mathfrak{C})$ in the following way:

$$\xi(I_A) \;=\; I_{\Phi(A)} \qquad \text{for every } A \in \mathcal{A} \tag{2.30}$$

$$\xi\Bigg(\sum_{j=1}^{m} a_j I_{A_j}\Bigg) \;=\; \sum_{j=1}^{m} a_j I_{\Phi(A_j)} \qquad \text{for every simple function on } (\Omega, \mathcal{A}) \tag{2.31}$$

and

$$\xi(f) \;=\; \lim_{n\to\infty} \xi(s_n) \qquad \text{for every } f \in \mathcal{L}_\infty(\Omega, \mathcal{A}) \tag{2.32}$$

where each s_n is a simple function and $\|s_n - f\| \to 0$.

Proposition 2.41 *Equations (2.30), (2.31) and (2.32) define a map*

$$\xi \;:\; \mathcal{L}_\infty(\Omega, \mathcal{A}) \;\to\; \mathcal{L}_\infty(\Xi, \mathfrak{C})\,, \qquad g \;\mapsto\; \xi(g) \tag{2.33}$$

which is an M-space isomorphism.

Proof: According to Lemma 8.28, where $\Psi = \Phi$, $(\mathcal{Y}, \mathcal{B}) = (\Omega, \mathcal{A})$ and $(\mathcal{Z}, \mathcal{D}) = (\Xi, \mathfrak{C})$, equations (2.30), (2.31) and (2.32) define a map

$$\xi \;:\; \mathcal{L}_\infty(\Omega, \mathcal{A}) \;\to\; \mathcal{L}_\infty(\Xi, \mathfrak{C})\,, \qquad g \;\mapsto\; \xi(g)$$

which is linear, positive and normalized – especially, ξ is norm-continuous.

Another application of Lemma 8.28, where $\Psi = \Phi^{-1}$, $(\mathcal{Y}, \mathcal{B}) = (\Xi, \mathfrak{C})$ and $(\mathcal{Z}, \mathcal{D}) = (\Omega, \mathcal{A})$, leads to a map

$$\varphi \;:\; \mathcal{L}_\infty(\Xi, \mathfrak{C}) \;\to\; \mathcal{L}_\infty(\Omega, \mathcal{A})\,, \qquad h \;\mapsto\; \varphi(h)$$

[11] For example, the clopen sets in $\mathbb{R}$ are $\emptyset$ and $\mathbb{R}$.

which is linear, positive and normalized (and, therefore, norm-continuous).

Obviously

$$\varphi\circ\xi(s)=s \qquad\text{and}\qquad \xi\circ\varphi(t)=t$$

for every simple function s on $(\Omega,\mathcal{A})$ and every simple function t on $(\Xi,\mathfrak{C})$. Hence, norm-continuity of ξ and φ implies that

$$\varphi\circ\xi(f)=f \qquad\text{and}\qquad \xi\circ\varphi(h)=h$$

for every $f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$ and every $h\in\mathcal{L}_\infty(\Xi,\mathfrak{C})$. That is, φ is the inverse of ξ – especially, ξ is bijective.

Next, for every $f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$

$$f\;\ge\;0 \qquad\Rightarrow\qquad \xi(f)\;\ge\;0$$

and

$$\xi(f)\;\ge\;0 \qquad\Rightarrow\qquad f\;=\;\varphi\big(\xi(f)\big)\;\ge\;0$$

Therefore, ξ is a vector lattice isomorphism according to Proposition 8.20. The properties of Φ imply

$$\|\xi(s)\|\;=\;\sup_{x\in\Xi}|\xi(s)(x)|\;=\;\sup_{x\in\Xi}|s(x)|\;=\;\|s\|$$

for every simple function s on $(\Omega,\mathcal{A})$. Hence, it follows from norm-continuity that

$$\|\xi(f)\|\;=\;\|f\| \qquad \forall\,f\in\mathcal{L}_\infty(\Omega,\mathcal{A})$$

□

Φ and ξ also induce a map

$$\phi\;:\;\;\mathrm{ba}(\Omega,\mathcal{A})\;\to\;\mathrm{ba}(\Xi,\mathfrak{C})$$

via $\phi(\mu)[h]\;=\;\mu\big[\xi^{-1}(h)\big]$ where ξ^{-1} is the inverse function of ξ . Since $\mathrm{ba}(\Omega,\mathcal{A})$ is the dual space of $\mathcal{L}_\infty(\Omega,\mathcal{A})$ and $\mathrm{ba}(\Xi,\mathfrak{C})$ is the dual space of $\mathcal{L}_\infty(\Xi,\mathfrak{C})$, ϕ is the adjoint operator of ξ^{-1}. It is easy to see that ϕ is

- linear
- positive: $\phi(\mu)\;\ge\;0 \qquad \forall\,\mu\ge 0$
- normalized: $\phi(\mu)[I_\Xi]\;=\;1 \qquad \forall\,\mu\ge 0$

Such maps are called *(generalized) randomizations* or *transitions* and will frequently be used later on; cf. Section 3.3.

ϕ has a nice continuity property which will become important in the following subsection where coherent upper previsions are represented by upper expectations.

Proposition 2.42 *Let*

$$\phi\;:\;\;\mathrm{ba}(\Omega,\mathcal{A})\;\to\;\mathrm{ba}(\Xi,\mathfrak{C})$$

be the adjoint operator of ξ^{-1} *as defined in the previous paragraphs. Endow* $\mathrm{ba}(\Omega,\mathcal{A})$ *with the* $\mathcal{L}_\infty(\Omega,\mathcal{A})$ *- topology and* $\mathrm{ba}(\Xi,\mathfrak{C})$ *with the* $\mathcal{L}_\infty(\Xi,\mathfrak{C})$ *- topology. Then,* ϕ *is continuous with respect to these topologies.*

Proof: All topological terms within this proof are with respect to the topologies mentioned in Proposition 2.42.

Let $(\mu_\gamma)_{\gamma\in D}$ be a net in $\mathrm{ba}(\Omega,\mathcal{A})$ which converges to some $\mu\in\mathrm{ba}(\Omega,\mathcal{A})$. This implies that, for every $h\in\mathcal{L}_\infty(\Xi,\mathfrak{C})$,

$$\phi(\mu_\gamma)[h] \;=\; \mu_\gamma\big[\xi^{-1}(h)\big] \;\underset{\gamma}{\longrightarrow}\; \mu\big[\xi^{-1}(h)\big] \;=\; \phi(\mu)[h]$$

That is, $\phi(\mu_\gamma) \underset{\gamma}{\longrightarrow} \phi(\mu)$ according to Theorem 8.24 b). □

Canonical Stone representation

This subsection ends with the description of a concrete space Ξ and a concrete map $\Phi:\mathcal{A}\to\mathfrak{B}$ in the Stone representation theorem (Theorem 2.40). This concrete description is not really important in this book. However, it is important that it is possible to uniquely determine a concrete choice of Ξ, Φ, ξ and ϕ. Later on, we will refer to this specific choice as "canonical Stone space", "canonical Stone isomorphism", "canonical Stone kernel" and "canonical Stone transition" respectively.

To this end, let Ξ be the set of all algebra homomorphisms

$$x \;:\; \mathcal{A} \;\to\; \{\emptyset,\Omega\}$$

A map x is called algebra homomorphism if, for every $A_1,\,A_2,\,A\,\in\,\mathcal{A}$,

$$x(A_1\cap A_2) \;=\; x(A_1)\,\cap\,x(A_2),\;\; x(A_1\cup A_2) \;=\; x(A_1)\,\cup\,x(A_2),\;\; x(A^{\mathrm{C}}) = \big(x(A)\big)^{\mathrm{C}}$$

Put

$$\Phi(A) \;=\; \big\{x\in\Xi \;\big|\; x(A)=\Omega\big\} \qquad \forall\,A\in\mathcal{A}$$

and endow Ξ with the topology generated by the base

$$\big\{\,\Phi(A) \;\big|\; A\in\mathcal{A}\,\big\}$$

Let $\mathfrak{C}$ denote the algebra of all clopen sets in Ξ. Then Ξ is a totally disconnected compact Hausdorff space and $\Phi:\mathcal{A}\to\mathfrak{C}$ is a bijective algebra homomorphism according to (Dunford and Schwartz, 1958, § I.12). That is, Ξ and Φ have all of the properties which are required in the Stone representation theorem (Theorem 2.40).

With this choice, Ξ is called *canonical Stone space of* $(\Omega,\mathcal{A})$ and Φ is called *canonical Stone isomorphism of* $(\Omega,\mathcal{A})$. According to Proposition 2.41, Φ induces an M-space isomorphism

$$\xi \;:\; \mathcal{L}_\infty(\Xi,\mathfrak{C}) \;\to\; \mathcal{L}_\infty(\Omega,\mathcal{A})\,, \qquad h \;\mapsto\; \xi(h)$$

and an adjoint operator of ξ^{-1},

$$\phi \;:\; \mathrm{ba}(\Omega,\mathcal{A}) \;\to\; \mathrm{ba}(\Xi,\mathfrak{C})$$

The map ξ is called *canonical Stone kernel of* $(\Omega,\mathcal{A})$ and ϕ is called *canonical Stone transition between* $(\Omega,\mathcal{A})$ *and* $(\Xi,\mathfrak{C})$ in this case.

2.5.3 Coherent upper previsions represented by upper expectations

This subsection investigates how an coherent upper prevision on $(\Omega, \mathcal{A})$ can be represented by an upper expectation. Since such a representation is possible for every coherent upper prevision, upper expectations are as general as coherent upper previsions from a theoretical point of view. The representation is based on the Stone representation theorem presented in Subsection 2.5.2. In particular, this means that coherent upper previsions on $(\Omega, \mathcal{A})$ are represented by upper expectations on the canonical Stone space Ξ which belongs to $(\Omega, \mathcal{A})$. In Subsection 2.5.2, we did not consider a σ-algebra but an algebra on Ξ. However, for the definition of upper expectations, we need a σ-algebra. Since the canonical Stone space Ξ is a compact Hausdorff space, we are in the situation of Subsection 2.4.4 and the Baire-σ-algebra turns out to be an appropriate choice.

Let Ω be a set and $\mathcal{A}$ an algebra on Ω. Let Ξ be the canonical Stone space of $(\Omega, \mathcal{A})$ and $\mathfrak{B}_0$ the Baire-σ-algebra on Ξ ; let $\mathfrak{C}$ denote the algebra of all clopen sets $C \subset \Xi$.

A clopen set C has the characteristic property that C and its complement C^{C} are open. Therefore, the indicator function I_C of a clopen set C is continuous! So, it follows from the definition of the Baire-σ-algebra that

$$\mathfrak{C} \subset \mathfrak{B}_0$$

In fact, the properties of Ξ imply that $\mathfrak{B}_0$ is the smallest σ-algebra on Ξ which contains $\mathfrak{C}$ and that

$$\mathcal{C}(\Xi) = \mathcal{L}_\infty(\Xi, \mathfrak{C}) \tag{2.34}$$

where $\mathcal{C}(\Xi)$ is the set of all continuous functions $f : \Xi \to \mathbb{R}$; cf. (Bhaskara Rao and Bhaskara Rao, 1983, p. 17f and Corollary 4.7.6(i)).

The following theorem is only a variant of a family of similar, well known theorems; cf. e.g. (Dunford and Schwartz, 1958, Lemma IV.9.11).

Theorem 2.43 *Let Ω be a set and $\mathcal{A}$ an algebra on Ω. Let Ξ be the canonical Stone space of $(\Omega, \mathcal{A})$ and $\mathfrak{B}_0$ the Baire-σ-algebra on Ξ. Then, there is a unique map*

$$\phi_0 : \mathrm{ba}(\Omega, \mathcal{A}) \to \mathrm{ca}(\Xi, \mathfrak{B}_0)$$

such that

$$\phi_0(\mu)(C) = \mu\big(\Phi^{-1}(C)\big) \qquad \forall\, C \in \mathfrak{C}, \quad \forall\, \mu \in \mathrm{ba}(\Omega, \mathcal{A}) \tag{2.35}$$

where $\Phi : \mathcal{A} \to \mathfrak{C}$ is the canonical Stone isomorphism.
Furthermore, ϕ_0 is an L-space isomorphism and

$$\phi_0(\mu)\big[\xi(f)\big] = \mu\big[f\big] \qquad \forall\, f \in \mathcal{L}_\infty(\Omega, \mathcal{A})\,, \;\; \forall\, \mu \in \mathrm{ba}(\Omega, \mathcal{A}) \tag{2.36}$$

where ξ is the canonical Stone kernel.

Proof:

[1] $$T_{\nu_0} : \ \mathcal{C}(\Xi) \ \to \ \mathbb{R}\,, \qquad f \ \mapsto \ \int f \, d\nu_0$$

is a (norm-)continuous linear functional for every $\nu_0 \in \mathrm{ca}(\Xi, \mathfrak{B}_0)$; cf. Theorem 2.34. Furthermore,

$$\varphi \ : \ \mathrm{ca}(\Xi, \mathfrak{B}_0) \ \to \ \mathcal{C}(\Xi)^*\,, \qquad \nu_0 \ \mapsto \ T_{\nu_0}$$

is an L-space isomorphism between $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ and the dual space of $\mathcal{C}(\Xi)$ according to Remark 2.35. Because of

$$\mathrm{ba}(\Xi, \mathfrak{C}) \ = \ \big(\mathcal{L}_\infty(\Xi, \mathfrak{C})\big)^* \ \overset{(2.34)}{=} \ \mathcal{C}(\Xi)^*$$

φ is an L-space isomorphism between $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ and $\mathrm{ba}(\Xi, \mathfrak{C})$ such that

$$\varphi(\nu_0)[h] \ = \ T_{\nu_0}[h] \ = \ \nu_0[h] \qquad \forall\, h \in \mathcal{C}(\Xi)\,, \quad \forall \nu_0 \in \mathrm{ca}(\Xi, \mathfrak{B}_0) \tag{2.37}$$

[2] Let ξ be the canonical Stone kernel; the inverse $\xi^{-1} : \mathcal{L}_\infty(\Xi, \mathfrak{C}) \ \to \ \mathcal{L}_\infty(\Omega, \mathcal{A})$ is an M-space isomorphism according to Proposition 2.41. Let $\phi = (\xi^{-1})^*$ be the adjoint operator of ξ^{-1}. Then, according to Proposition 8.22,

$$\phi \ : \ \ \mathrm{ba}(\Omega, \mathcal{A}) \ \to \ \mathrm{ba}(\Xi, \mathfrak{C})$$

is an L-space isomorphism.

[3] That is,

$$\mathrm{ba}(\Omega, \mathcal{A}) \ \xrightarrow{\phi} \ \mathrm{ba}(\Xi, \mathfrak{C}) \ \xrightarrow{\varphi^{-1}} \ \mathrm{ca}(\Xi, \mathfrak{B}_0)$$

where ϕ and φ^{-1} are L-space isomorphisms. Then,

$$\phi_0 \ := \ \varphi^{-1} \circ \phi \tag{2.38}$$

is an L-space isomorphism such that, for every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$ and $\mu \in \mathrm{ba}(\Omega, \mathcal{A})$,

$$\begin{aligned} \phi_0(\mu)[\xi(f)] \ &= \ \varphi^{-1}\Big(\phi(\mu)\Big)[\xi(f)] \ \overset{(2.37),(2.34)}{=} \ \phi(\mu)[\xi(f)] \ = \\ &= \ \mu\big[\xi^{-1}\big(\xi(f)\big)\big] \ = \ \mu[f] \end{aligned}$$

That is, ϕ_0 fulfills (2.36). Especially, ϕ_0 fulfills (2.35) – to see this, put $f \ = \ I_{\Phi^{-1}(C)}$ for $C \in \mathfrak{C}$ and note that $I_C \ = \ \xi(f)\,.$

[4] Now, let $\sigma : \mathrm{ba}(\Omega, \mathcal{A}) \to \mathrm{ca}(\Xi, \mathfrak{B}_0)$ be another map which fulfills (2.35). Take any $\mu \in \mathrm{ba}(\Omega, \mathcal{A})$. Then,

$$\sigma(\mu)(C) \ \overset{(2.35)}{=} \ \mu\big(\Phi^{-1}(C)\big) \ \overset{(2.35)}{=} \ \phi_0(\mu)(C) \qquad \forall\, C \in \mathfrak{C}$$

and it follows that $\sigma(\mu) = \phi_0(\mu)$ because $\mathfrak{C}$ is a $\cap$-stable generator of $\mathfrak{B}_0$.

□

Especially, Theorem 2.43 says that $\mathrm{ba}(\Omega, \mathcal{A})$ and $\mathrm{ca}(\Xi, \mathfrak{B}_0)$ are L-space isomorphic. The uniquely determined map ϕ_0 in Theorem 2.43 is called *canonical Stone transition between* $(\Omega, \mathcal{A})$ *and* $(\Xi, \mathfrak{B}_0)$.

Now, the main theorem of this section can be formulated:

Theorem 2.44 (Canonical Stone representation) *Let Ω be a set and $\mathcal{A}$ an algebra on Ω. Let Ξ be the canonical Stone space, ξ the canonical Stone kernel and ϕ_0 the canonical Stone transition between $(\Omega, \mathcal{A})$ and $(\Xi, \mathfrak{B}_0)$.*
Then, for every coherent upper prevision $\overline{P}$ with credal set $\mathcal{M}$ on $(\Omega, \mathcal{A})$,

$$\mathcal{P}_0 \;:=\; \phi_0(\mathcal{M})$$

is the structure of an upper expectation $\overline{P}_0$ on $(\Xi, \mathfrak{B}_0)$ such that

$$\overline{P}[f] \;=\; \overline{P}_0\big[\xi(f)\big] \qquad \forall\, f \in \mathcal{L}_\infty(\Omega, \mathcal{A}) \tag{2.39}$$

Furthermore, the structure $\mathcal{P}_0$ is $\mathcal{C}(\Xi)$ - compact and can be written as

$$\mathcal{P}_0 \;=\; \big\{P_0 \in \mathrm{ca}_1^+(\Xi, \mathfrak{B}_0) \;\big|\; P_0[h] \le \overline{P}_0[h] \;\; \forall\, h \in \mathcal{C}(\Xi)\big\} \tag{2.40}$$

$\overline{P}_0$ is called canonical Stone representation of *$\overline{P}$.*

Proof:

[1] Put $\overline{P}_0[h_0] \;=\; \sup_{P_0 \in \mathcal{P}_0} P_0[h_0]$ for every $h_0 \in \mathcal{L}_\infty(\Xi, \mathfrak{B}_0)$.

[2] At first, (2.39) is shown: For every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$,

$$\overline{P}_0\big[\xi(f)\big] \;=\; \sup_{P_0 \in \mathcal{P}_0} P_0\big[\xi(f)\big] \;=\; \sup_{P \in \mathcal{M}} \phi_0(P)\big[\xi(f)\big] \;\overset{(2.36)}{=}\; \sup_{P \in \mathcal{M}} P[f] \;=\; \overline{P}[f]$$

[3] Next, (2.40) is shown:
The definition of $\overline{P}_0$ implies "$\subset$" in (2.40). For the proof of "$\supset$", take any $P_0 \in \mathrm{ca}_1^+(\Xi, \mathfrak{B}_0)$ such that

$$P_0[h] \;\le\; \overline{P}_0[h] \qquad \forall\, h \in \mathcal{C}(\Xi)$$

and put $P \;:=\; \phi_0^{-1}(P_0)$. Then, for every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$,

$$P[f] \;\overset{(2.36)}{=}\; \phi_0(P)\big[\xi(f)\big] \;=\; P_0\big[\xi(f)\big] \;\overset{(2.34)}{\le}\; \overline{P}_0\big[\xi(f)\big] \;\overset{(2.39)}{=}\; \overline{P}[f]$$

for every $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$. That is, $P \in \mathcal{M}$ and, therefore, $P_0 = \phi_0(P) \in \mathcal{P}_0$.

[4] Especially, (2.40) implies that $\mathcal{P}_0$ is the structure of an upper expectation.

[5] $\mathcal{C}(\Xi)$ - compactness of $\mathcal{P}_0$ follows from (2.40) according to Theorem 2.36.

□

That is, for every set Ω and every algebra $\mathcal{A}$ on Ω, there is a compact Hausdorff space such that every coherent upper prevision on $(\Omega, \mathcal{A})$ can be represented by an upper expectation on $(\Xi, \mathfrak{B}_0)$ whose structure is $\mathcal{C}(\Xi)$ - compact. Therefore, the setup in Subsection 2.4.4 – namely upper expectations on compact Hausdorff spaces and the $\mathcal{C}(\Xi)$ - topology – is as general as the setup in Section 2.3 (coherent upper previsions and $\mathcal{L}_\infty(\Omega, \mathcal{A})$). If we are faced with a coherent upper prevision $\overline{P}$ on $(\Omega, \mathcal{A})$ whose credal set $\mathcal{M}$ contains probability charges which are no probability measures, we can always turn over to the classical measure theoretic setup in the following way: Build the set

$$\mathcal{P} \;:=\; \big\{\phi_0(P) \;\big|\; P \in \mathcal{M}\big\} \;=\; \phi_0(\mathcal{M})$$

This is a $\mathcal{C}(\Xi)$ - compact structure of an upper expectation which consists of probability measures on $(\Xi, \mathfrak{B}_0)$ where Ξ is a compact Hausdorff space. Then, all calculations can be done on the measurable space $(\Xi, \mathfrak{B}_0)$ and with probability measures. Finally, the obtained results can be transformed back to $(\Omega, \mathcal{A})$ by

$$\overline{P}[f] \;=\; \overline{P}_0\big[\xi(f)\big] \qquad \forall\, f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$$

where $\overline{P}_0$ is the upper expectation which belongs to the structure $\mathcal{P}$ on $(\Xi, \mathfrak{B}_0)$.

The reader who does not like probability charges which are not probability measures may always turn over to the classical measure theoretic setup by this way. However, the canonical Stone space is not very convenient to handle so that it seems to be easier to deal with coherent upper previsions on $(\Omega, \mathcal{A})$. Accordingly, the practical use of this representation may be rather limited but it is very interesting from a theoretical point of view that there is no essential difference between both setups. The rest of this book usually considers coherent upper previsions.

The specific choice of the canonical Stone representation is not important in this book. However, it is important that there is a uniquely defined representation – this will be used in Subsection 3.3.3 in order to define standard measures for probability charges which are not σ-additive. Standard measures are useful because it is possible to calculate (upper) Bayes risks in decision theory with the help of standard measures. These concepts of decision theory under complex uncertainty are presented in the following chapter. Therein, complex uncertainty is modeled by coherent upper previsions and credal sets.

Chapter 3

Extended decision theoretic framework

3.1 Introduction

Decision theory provides a formal framework for determining optimal actions under uncertainty on the states of nature. It has a wide range of potential areas of application which includes also statistical problems. However, a serious problem in practical applications of decision theory is that the uncertainty often is too complex to be adequately described by a precise probability distribution. As explained in Chapter 1, ambiguity is an important part of decision making which cannot be neglected. In order to take ambiguity into account properly, any of the concepts of imprecise probabilities presented in Chapter 2 can be used. Imprecise probabilities (or equivalent concepts) are already applied in many decision theoretic evaluations, for example in mathematical economics – e.g. Gilboa and Schmeidler (1989), Schied (2006), Maccheroni et al. (2006) and Föllmer et al. (2007) – and in articles concerning climate change – e.g. Kriegler (2005) and Hall et al. (2007).
A general article about decision making where uncertainties are modeled by coherent lower previsions is de Cooman and Walley (2002). Different optimality criteria are discussed by Schervish et al. (2003) and Troffaes (2007) in this setup. Algorithms for the calculation of optimal decisions are given by Kikuti et al. (2005) and Utkin and Augustin (2005).
Within this book, the concept of imprecise probabilities according to Walley (1991) – i.e. the concept of coherent upper previsions – is used. The present chapter introduces the decision theoretic setup and develops some important tools which prove to be useful in decision theory under imprecise probabilities. These tools are mainly adapted from the work of L. Le Cam (Le Cam, 1986) and transfered to the theory of imprecise probabilities. Within the theory of imprecise probabilities, these tools have been introduced in Hable (2007) and Hable (2008b).

We start with an informal description of the decision theoretic setup under imprecise probabilities. In order to explain the decision theoretic setup we are concerned with, the classical decision theoretic setup is recalled at first:

There is a set Θ where each element $\theta \in \Theta$ represents a possible state of nature. We know that one state of nature will occur but we do not know which one it will be. Furthermore, there is a set $\mathbb{D}$ where each element $t \in \mathbb{D}$ is a decision – also called action – we can choose. Depending on what state of nature θ occurs, every decision t leads to a loss $W_\theta(t) \in \mathbb{R}$. The goal is to choose a "good" decision so that the loss is as small as possible.

Sometimes, we might know a precise expectation π for the states of nature $\theta \in \Theta$. Then, we can choose the decision that minimizes the expected loss

$$\int_\Theta W_\theta(t)\,\pi(d\theta)$$

In addition, we often can choose our decision on base of an observation $y \in \mathcal{Y}$. For example, y may be the outcome of an experiment. The distribution of the observation y might be a precise expectation Q_θ which depends on the state of nature θ. That is $(Q_\theta)_{\theta\in\Theta}$ is a model which describes the distribution of the observation y.
Such "data-based decision making" can be formalized by choosing a decision function $\delta:\ \mathcal{Y} \to \mathbb{D}\,,\ \ y \mapsto \delta(y)$ which minimizes

$$\int_\Theta \int_\mathcal{Y} W_\theta(\delta(y))\,Q_\theta(dy)\,\pi(d\theta)$$

Decision theory commonly also deals with randomized decisions. Randomized decision procedures (randomizations) are defined in Section 3.2 and Subsection 3.3.1. Confer Berger (1985) for an introduction to these basic concepts of decision theory.

In the following, we are concerned with a more general decision theoretic setup because we also want to deal with imprecise probabilities:

Since the prior knowledge about the states of nature will frequently not be precise, we allow for a whole set $\mathcal{P}$ of possible precise expectations π. Also the knowledge about the distribution of the observation may only be imprecise so that there are sets $\mathcal{M}_\theta$ of possible precise expectations Q_θ. While minimizing the expected loss in case of precise expectations is widely accepted, there are several reasonable optimality criteria in case of imprecise expectations; confer Troffaes (2007) for a discussion of the most important ones. In this book, the so-called Γ-minimax criterion is mainly used which represents a worst case consideration.[1] That is we choose a decision function δ (or rather a randomization later on) which minimizes the twofold upper expectation

$$\sup_{\pi\in\mathcal{P}} \int_\Theta \sup_{Q_\theta\in\mathcal{M}_\theta} \int_\mathcal{Y} W_\theta(\delta(y))\,Q_\theta(dy)\,\pi(d\theta)$$

Unfortunately, a direct solution of this problem is quite often computationally intractable. This fact gives rise to many of the investigations in this book.
A solution of this problem is much more easier if decisions are not data-based. However, data-based decision problems are more important for applications – the more so as the main applications we are interested in are statistical problems. In fact, statistical problems can be formalized as decision theoretic problems. The part of decision theory which is concerned with the formalization of statistical problems is also called statistical decision theory (cf. e.g. Berger (1985) or Le Cam (1986)). Of course, statistical decision theory is always data-based.

Remark 3.1 *In case of precise probabilities, it is usually not necessary to explicitly consider data-based decision problems because it is possible to solve data-based decision problems by solving appropriate data-free decision problems. This is due to a famous theorem which is often called the "main theorem of Bayesian decision theory"; cf. e.g. (Berger, 1985, § 4.4.1). However, this is usually not possible in case of imprecise probabilities as pointed out by Augustin (2003).*[2] *As a consequence, data-based decision problems are*

[1] For the use of the Γ-minimax criterion in Bayesian analysis, cf. Vidakovic (2000) and the literature cited therein.

[2] In robust Bayesian analysis, this is also explained in Vidakovic (2000).

a matter of its own in imprecise probability theory. The question if data-based decision problems have to be considered explicitly picks up an old debate between frequentists and Bayesians: Does the posterior distribution $\pi(\cdot|y)$ contain all relevant information after observing y? It is neither the aim of the present book to add new arguments to this debate nor to review it. The present book is only concerned with the mathematical investigation of decision problems with an explicitly data-based formulation.
Decision problems under imprecise probabilities which are explicitly data-based have hardly been considered before. One of the very few exceptions is e.g. Augustin (2004).

The following Section 3.2 contains a mathematical rigorous explanation of the decision theoretic setup. Here, fundamental decision theoretic concepts are recalled and extended to imprecise probabilities.

Section 3.3 introduces some important advanced decision theoretic tools, namely generalized randomizations, equivalence/sufficiency and standard measures. As already mentioned above, many of these concepts are translations of objects which have been introduced by L. Le Cam in a very general setup of *precise* probabilities. Most of these translations are analogous to the proceeding in Buja (1984). However, there are fundamental differences which arise from the fact that Buja (1984) uses more traditional concepts based on Polish spaces and σ-additive probability measures.

Firstly, Subsection 3.3.1 is concerned with generalized randomizations which generalize Markov kernels. On the one hand, these generalized randomizations have a less descriptive interpretation but, on the other hand, they are a powerful mathematical tool. Indeed, they help to avoid some difficulties which arise if Markov kernels are considered only. In contrast to Le Cam, results which are obtained by use of generalized randomizations are translated in terms of Markov kernels as far as possible in this book.

In Subsection 3.3.2, we define an equivalence relation on the set of all (precise) models $(Q_\theta)_{\theta\in\Theta}$ according to which two (precise) models $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ are equivalent if the following is true: Observations of model $(P_\theta)_{\theta\in\Theta}$ can artificially be generated (by a randomization) from observations of model $(Q_\theta)_{\theta\in\Theta}$ and vice versa.

Every equivalence class contains a uniquely determined standard representative. This representative which is called *standard model* is defined in Subsection 3.3.3. A standard model is a model which consists of probability measures on a very convenient measurable space. Due to equivalence, we can investigate every model with the help of its standard model (cf. Secion 4.2). In order to define standard models for precise models which do not consist of σ-additive probability measures, the results from Section 2.5 concerning canonical Stone representations are crucial.

Since the concepts introduced in Section 3.3 are strongly connected with concepts introduced by L. Le Cam, Chapter 3 closes with Section 3.4 where the connections to L. Le Cam's setup are explained. On the one hand, L. Le Cam's setup is more specific than the setup used in imprecise probabilities because L. Le Cam only deals with precise probabilities. On the other hand, his setup is more general because he does not consider explicitly specified sample spaces but considers probabilities as elements of certain vector lattices. Furthermore, Section 3.4 may also serve as a comprehensible introduction to L. Le Cam's abstract setup.

3.2 Basic definitions in decision theory

Let Θ be any index set. The elements $\theta \in \Theta$ are called *states of nature* and Θ is called *set of states.*

Let $\mathbb{D}$ be any set. The elements $t \in \mathbb{D}$ are called *decisions.* Let $\mathcal{D}$ be an algebra on $\mathbb{D}$. Then, $(\mathbb{D}, \mathcal{D})$ is called *decision space.* $\mathbb{D}$ represents the set of all possible decisions in a decision problem.
A family of functions

$$(W_\theta)_{\theta\in\Theta} \;\subset\; \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$$

is called *loss function.* Every loss function $(W_\theta)_{\theta\in\Theta}$ defines a function

$$W \;:\; \Theta \times \mathcal{D} \;\to\; \mathbb{R}\,, \qquad (\theta, t) \;\mapsto\; W_\theta(t)$$

This function is also called *loss function.*

Let $\mathcal{Y}$ be a set and $\mathcal{B}$ an algebra on $\mathcal{Y}$. The elements of $\mathcal{Y}$ represent the possible outcomes of an experiment. Therefore, $(\mathcal{Y}, \mathcal{B})$ is called *sample space.*
A measurable map

$$\delta \;:\; \mathcal{Y} \;\to\; \mathbb{D}\,, \qquad y \;\mapsto\; \delta(y)$$

is called *decision function.* A finitely additive Markov kernel

$$\tau \;:\; \mathcal{Y} \times \mathcal{D} \;\to\; \mathbb{R}\,, \qquad (y, D) \;\mapsto\; \tau_y(D)$$

is called *randomized decision function* (on $(\mathcal{Y}, \mathcal{B})$); confer Subsection 3.3.1 for finitely additive Markov kernels. Especially, τ defines a map

$$\tau_\bullet \;:\; \mathcal{Y} \;\to\; \mathrm{ba}_1^+(\mathbb{D}, \mathcal{D})\,, \qquad y \;\mapsto\; \tau_y$$

$y \;\mapsto\; \tau_y$ has the following descriptive interpretation: After observing y, start an auxiliary random experiment according to the distribution τ_y and choose that action d which is the outcome of the auxiliary random experiment.

A family of probability charges

$$(Q_\theta)_{\theta\in\Theta} \;\subset\; \mathrm{ba}_1^+(\mathcal{Y}, \mathcal{B})$$

is called *precise model on* $(\mathcal{Y}, \mathcal{B})$.
A family of coherent upper previsions

$$(\overline{Q}_\theta)_{\theta\in\Theta} \quad \text{on} \quad \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$$

is called *imprecise model on* $(\mathcal{Y}, \mathcal{B})$.
These terms are adopted from the notion “statistical model”. Buja (1984) and Le Cam (1964), for example, use the term “experiment” instead of “model”.
Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y}, \mathcal{B})$ and, for every $\theta \in \Theta$, let $\mathcal{M}_\theta$ be the credal set of $\overline{Q}_\theta$ on $(\mathcal{Y}, \mathcal{B})$. Then, the family of credal sets

$$(\mathcal{M}_\theta)_{\theta\in\Theta}\,, \qquad \mathcal{M}_\theta \;\subset\; \mathrm{ba}_1^+(\mathcal{Y}, \mathcal{B})$$

is denoted as the family of credal sets which corresponds to the imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$.

Notation 3.2 $\big(\mathcal{Y},\mathcal{B},(Q_\theta)_{\theta\in\Theta}\big)$ *is called* precise model *if* $\mathcal{Y}$ *is a set with algebra* $\mathcal{B}$ *and* $(Q_\theta)_{\theta\in\Theta}$ *is a precise model on* $(\mathcal{Y},\mathcal{B})$ *.*
$\big(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta}\big)$ *is called* imprecise model *if* $\mathcal{Y}$ *is a set with algebra* $\mathcal{B}$ *and* $(\overline{Q}_\theta)_{\theta\in\Theta}$ *is an imprecise model on* $(\mathcal{Y},\mathcal{B})$ *.*

With all these settings, the function

$$\Theta \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; Q_\theta\Big[\tau_\bullet[W_\theta]\Big] \;=\; \int_{\mathcal{Y}}\int_{\mathbb{D}} W_\theta(t)\,\tau_y(dt)Q_\theta(dy)$$

is called *risk function of* τ (for the precise model $(Q_\theta)_{\theta\in\Theta}$); and the function

$$\Theta \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; \overline{Q}_\theta\Big[\tau_\bullet[W_\theta]\Big] \;=\; \sup_{Q_\theta\in\mathcal{M}_\theta}\int_{\mathcal{Y}}\int_{\mathbb{D}} W_\theta(t)\,\tau_y(dt)Q_\theta(dy)$$

is called *risk function of* τ (for the imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$).

The lower the risk function is the better the (randomized) decision function is. Clearly, a (randomized) decision function $\tilde{\tau}$ is optimal if

$$\overline{Q}_\theta\Big[\tilde{\tau}_\bullet[W_\theta]\Big] \;\le\; \overline{Q}_\theta\Big[\tau_\bullet[W_\theta]\Big] \qquad \forall\,\theta\in\Theta$$

for every other randomized decision function τ. Unfortunately, such a "uniformly optimal" $\tilde{\tau}$ almost never exists. Therefore, we have to rely on different optimality criteria defined by the Bayes risk:
Let π be a probability charge on $(\Theta,2^\Theta)$. Then, π is called *(precise) prior distribution on* Θ and

$$\begin{aligned} R_\pi\big((Q_\theta)_{\theta\in\Theta},\tau,W\big) \;&=\; \int_\Theta Q_\theta\Big[\tau_\bullet[W_\theta]\Big]\,\pi(d\theta) \;=\\ &=\; \int_\Theta\int_{\mathcal{Y}}\int_{\mathbb{D}} W_\theta(t)\,\tau_y(dt)Q_\theta(dy)\,\pi(d\theta)\end{aligned}$$

is called *Bayes risk of* τ *with respect to* π.
More generally, let $\overline{\Pi}$ be an coherent upper prevision on $\mathcal{L}_\infty(\Theta,2^\Theta)$ with corresponding credal set $\mathcal{P}$ and consider an imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ with corresponding family of credal sets $(\mathcal{M}_\theta)_{\theta\in\Theta}$. Then, $\overline{\Pi}$ is called *imprecise prior distribution on* Θ and

$$\begin{aligned} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\tau,W\big) \;&=\; \sup_{\pi\in\mathcal{P}}\int_\Theta \overline{Q}_\theta\Big[\tau_\bullet[W_\theta]\Big]\,\pi(d\theta) \;=\\ &=\; \sup_{\pi\in\mathcal{P}}\int_\Theta \sup_{Q_\theta\in\mathcal{M}_\theta}\int_{\mathcal{Y}}\int_{\mathbb{D}} W_\theta(t)\,\tau_y(dt)Q_\theta(dy)\,\pi(d\theta)\end{aligned}$$

is called *(upper) Bayes risk of* τ *with respect to* $\overline{\Pi}$.

A (randomized) decision function $\tilde{\tau}$ is called optimal with respect to the prior $\overline{\Pi}$ if

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\tilde{\tau},W\big) \;\le\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\tau,W\big)$$

for every other randomized decision function τ. That is, $\tilde{\tau}$ is optimal if it minimizes the upper Bayes risk.

These definitions includes that we have chosen the Γ-minimax optimality criterion which represents a worst case consideration (cf. Section 3.1) - as done e.g. in Huber and Strassen (1973) and Buja (1984) in a similar setup or in robust Bayesian analysis (c.f. Vidakovic (2000)).

Notation:

It will be seen in the following section that every randomized decision function τ defines a map

$$\mathrm{ba}(\mathcal{Y},\mathcal{B}) \;\to\; \mathrm{ba}(\mathbb{D},\mathcal{D})$$

where the image of $\mu \in \mathrm{ba}(\mathcal{Y},\mathcal{B})$ is the bounded charge on $\mathrm{ba}(\mathbb{D},\mathcal{D})$ given by

$$\mathcal{L}_\infty(\mathbb{D},\mathcal{D}) \;\to\; \mathbb{R}\,, \qquad h \;\mapsto\; \mu\Big[\tau_\bullet[h]\Big] \;=\; \int_{\mathcal{Y}} h(t)\,\tau_y(dt)\,\mu(dy)$$

This map is again denoted by τ. That is,

$$\tau(\mu)[h] \;=\; \mu\Big[\tau_\bullet[h]\Big] \;=\; \int_{\mathcal{Y}} h(t)\,\tau_y(dt)\,\mu(dy)$$

With this notation, the risk function of τ can be written as

$$\Theta \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; \tau(Q_\theta)[W_\theta]$$

3.3 Extended decision theoretic concepts

3.3.1 Generalized Randomizations

3.3.1.1 Definitions and basic properties

Usually, randomizations are modeled via Markov kernels. Since σ-additivity is relaxed to finite additivity in this book, it is suggesting to model randomizations via "finitely additive Markov kernels".

Definition 3.3 *Let Ω_1 be a set with algebra $\mathcal{A}_1$ and let Ω_2 be another set with algebra $\mathcal{A}_2$. A* finitely additive Markov kernel on $\Omega_1 \times \mathcal{A}_2$ *is a map*

$$\tau \;:\; \Omega_1 \times \mathcal{A}_2 \;\to\; \mathbb{R}\,, \qquad (\omega_1, A_2) \;\mapsto\; \tau_{\omega_1}(A_2)$$

such that

- $\tau_\bullet(A_2) :\ \omega_1 \mapsto \tau_{\omega_1}(A_2)$ *is an element of* $\mathcal{L}_\infty(\Omega_1,\mathcal{A}_1)$ *for every* $A_2 \in \mathcal{A}_2$ *and*
- $\tau_{\omega_1} :\ A_2 \mapsto \tau_{\omega_1}(A_2)$ *is an element of* $\mathrm{ba}_1^+(\Omega_2,\mathcal{A}_2)$ *for every* $\omega_1 \in \Omega_1$.

A finitely additive Markov kernel on $\Omega_1 \times \mathcal{A}_2$ is also called randomized function from $(\Omega_1,\mathcal{A}_1)$ to $(\Omega_2,\mathcal{A}_2)$.

The only difference between this definition and the usual definition of a Markov kernel is: Here, we do not insist on $\tau_{\omega_1} \in \mathrm{ca}_1^+(\Omega_2,\mathcal{A}_2)$ – we only insist on $\tau_{\omega_1} \in \mathrm{ba}_1^+(\Omega_2,\mathcal{A}_2)$. This explains the term "finitely additive Markov kernel".

A *non-randomized* function from $(\Omega_1, \mathcal{A}_1)$ to $(\Omega_2, \mathcal{A}_2)$ is a measurable function $\delta : \Omega_1 \to \Omega_2$ which maps a fixed $\omega_1 \in \Omega_1$ to a fixed $\omega_2 = \delta(\omega_1) \in \Omega_2$. That is, every ω_1 leads to some $\omega_2 = \delta(\omega_1)$ in a deterministic way.
The idea behind a *randomized* function from $(\Omega_1, \mathcal{A}_1)$ to $(\Omega_2, \mathcal{A}_2)$ is the following procedure: Given some ω_1, start a auxiliary random experiment according to the distribution τ_{ω_1}. Then, this auxiliary random experiment produces the ω_2 in a random way.
Finitely additive Markov kernels are called *ordinary* randomizations because they are – apart from σ-additivity – exactly the randomizations which are usually used in decision theory and because they have a descriptive interpretation as randomized functions. Below, a slight generalization will be defined which is called *generalized randomizations*.

Firstly, note that a finitely additive Markov kernel τ defines a map

$$T \;:\; \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \;\to\; \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)\,, \qquad f_2 \;\mapsto\; T(f_2)$$

via

$$T(f_2)(\omega_1) \;=\; \int_{\Omega_2} f_2(\omega_2)\, \tau_{\omega_1}(d\omega_2) \tag{3.1}$$

for every $\omega_1 \in \Omega_1$ and $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$. This map $T : \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \to \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$ is

- linear
- positive: $T(f_2) \;\geq\; 0 \quad \forall\, f_2 \;\geq\; 0$
- normalized: $T(I_{\Omega_2}) \;=\; I_{\Omega_1}$

Furthermore, a finitely additive Markov kernel τ defines a map

$$\sigma \;:\; \mathrm{ba}(\Omega_1, \mathcal{A}_1) \;\to\; \mathrm{ba}(\Omega_2, \mathcal{A}_2)\,, \qquad \sigma \;\mapsto\; \sigma(\mu_1)$$

via

$$\sigma(\mu_1)[f_2] \;=\; \int_{\Omega_2} f_2(\omega_2)\, \tau_{\omega_1}(d\omega_2)\, \mu_1(d\omega_1) \tag{3.2}$$

for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$. This map $\sigma : \mathrm{ba}(\Omega_1, \mathcal{A}_1) \to \mathrm{ba}(\Omega_2, \mathcal{A}_2)$, is

- linear
- positive: $\sigma(\mu_1) \;\geq\; 0 \quad \forall\, \mu_1 \;\geq\; 0$
- normalized: $\sigma(\mu_1)[I_{\Omega_2}] \;=\; \mu_1[I_{\Omega_2}] \quad \forall\, \mu_1$

Note, that σ is the adjoint operator of T because

$$\sigma(\mu_1)[f_2] \;=\; \mu_1\big[T(f_2)\big] \qquad \forall\, f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)\,, \quad \forall\, \mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$$

As in (Le Cam, 1964, § 3) and (Le Cam, 1986, § 1.3), this motivates the following definition:

Definition 3.4 (Generalized randomization) *Let Ω_1 be a set with algebra $\mathcal{A}_1$ and let Ω_2 be another set with algebra $\mathcal{A}_2$. A* generalized randomization from $(\Omega_1,\mathcal{A}_1)$ to $(\Omega_2,\mathcal{A}_2)$ *is a map*

$$\sigma \;:\; \mathrm{ba}(\Omega_1,\mathcal{A}_1) \;\to\; \mathrm{ba}(\Omega_2,\mathcal{A}_2)\,, \qquad \sigma \;\mapsto\; \sigma(\mu_1)$$

which is

- *linear*
- *positive:* $\sigma(\mu_1) \;\geq\; 0 \quad \forall\,\mu_1 \;\geq\; 0\,,\quad \mu_1 \in \mathrm{ba}(\Omega_1,\mathcal{A}_1)$
- *normalized:* $\sigma(\mu_1)[I_{\Omega_2}] \;=\; \mu_1[I_{\Omega_2}] \quad \forall\,\mu_1 \in \mathrm{ba}(\Omega_1,\mathcal{A}_1)$

$\mathcal{T}(\Omega_1,\Omega_2)$ *denotes the set of all generalized randomizations from* $(\Omega_1,\mathcal{A}_1)$ *to* $(\Omega_2,\mathcal{A}_2)$.

Remark 3.5 *The above definition is a translation of the definitions of "randomization" in (Le Cam, 1964, § 3) and "transition" in (Le Cam, 1986, § 1.3). Due to the usual setup based on explicitly specified sample spaces $(\Omega_i,\mathcal{A}_i)$, domain and codomain of generalized randomizations are* $\mathrm{ba}(\Omega_1,\mathcal{A}_1)$ *and* $\mathrm{ba}(\Omega_2,\mathcal{A}_2)$ *in Definition 3.4. In contrast, the definition of transitions in (Le Cam, 1986, § 1.3) is formulated in terms of general L-spaces – this is due to the general setup in Le Cam (1986) where the sample spaces are not explicitly specified. The definition of transitions is recalled in Section 3.4 and Proposition 3.36 shows that every generalized randomization in the sense of Definition 3.4 is a transition in the sense of (Le Cam, 1986, § 1.3).*

As seen above, every (finitely additive) Markov kernel defines a generalized randomization. Since those generalized randomizations which are defined by (finitely additive) Markov kernels are exactly the objects which are usually considered as randomizations, we may call them ordinary randomizations:

Definition 3.6 (Ordinary randomization)
A generalized randomization (from $(\Omega_1,\mathcal{A}_1)$ to $(\Omega_2,\mathcal{A}_2)$) which is defined by a finitely additive Markov kernel via (3.2) is called ordinary randomization (from $(\Omega_1,\mathcal{A}_1)$ to $(\Omega_2,\mathcal{A}_2)$) *or simply* randomization.
$\mathcal{T}_0(\Omega_1,\Omega_2)$ *denotes the set of all (ordinary) randomizations from* $(\Omega_1,\mathcal{A}_1)$ *to* $(\Omega_2,\mathcal{A}_2)$.

Of course, every ordinary randomization is a generalized randomization but even more: The ordinary randomizations are dense in the set of the generalized randomizations; cf. Theorem 3.10. Later on, we will also need a class of randomizations which have a very simple form; those randomizations are called *restricted randomizations*:

Definition 3.7 (Restricted randomization) *For $i \in \{1,2\}$, let Ω_i be a set with algebra $\mathcal{A}_i$ and let*

$$\tau \;:\; \Omega_1 \times \mathcal{A}_2 \;\to\; \mathbb{R}$$

be a finitely additive Markov kernel on $\Omega_1 \times \mathcal{A}_2$ such that

$$\tau(\omega_1,A_2) \;=\; \sum_{\tilde\omega_2\in\tilde\Omega_2} \alpha_{\tilde\omega_2}(\omega_1)\cdot\delta_{\tilde\omega_2}(A_2) \qquad \forall\,\omega_1\in\Omega_1\,,\quad A_2\in\mathcal{A}_2 \tag{3.3}$$

where $\tilde\Omega_2 \subset \Omega_2$ is a finite set, $\delta_{\tilde\omega_2}$ denotes the Dirac measure in $\tilde\omega_2$,

$$\alpha_{\tilde\omega_2} \geq 0\,,\quad \alpha_{\tilde\omega_2} \in \mathcal{L}_\infty(\Omega_2,\mathcal{A}_2) \quad \forall\,\tilde\omega_2 \in \tilde\Omega_2 \qquad \textit{and} \qquad \sum_{\tilde\omega_2\in\tilde\Omega_2} \alpha_{\tilde\omega_2} \equiv 1$$

Then, the ordinary randomization which is defined by τ via (3.2) is called restricted randomization (from $(\Omega_1, \mathcal{A}_1)$ to $(\Omega_2, \mathcal{A}_2)$) *and* $\mathcal{T}_r(\Omega_1, \Omega_1)$ *denotes the set of all restricted randomizations from* $(\Omega_1, \mathcal{A}_1)$ *to* $(\Omega_2, \mathcal{A}_2)$ *.*

Remark 3.8 *Analogously to the definition of ordinary randomizations, the above definition is a translation of the definitions of "restricted randomized map" in (Le Cam, 1964, § 3) and "finitely supported transition" in (Le Cam, 1986, § 1.4). According to Proposition 3.37, the restricted randomizations in the sense of Definition 3.7 are precisely the* $((\Gamma, H)-$*continuous) finitely supported transitions in the sense of (Le Cam, 1986, § 1.4).*

$\mathcal{T}(\Omega_1, \Omega_2)$ can be provided with the topology of pointwise convergence. This is the smallest topology so that

$$\mathcal{T}(\Omega_1, \Omega_2) \to \mathbb{R}, \qquad \sigma \mapsto \sigma(\mu_1)[f_2]$$

is continuous for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and every $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$. The following theorem is one of the reasons why we use this generalization of randomized functions:

Theorem 3.9 $\mathcal{T}(\Omega_1, \Omega_2)$ *is a compact Hausdorff space (with respect to the topology of pointwise convergence).*
(Cf. (Le Cam, 1986, Theorem 1.4.2).)

The following theorem indicates that the term "randomization" has only been slightly generalized:

Theorem 3.10 *The following inclusions are valid:*

$$\mathcal{T}_r(\Omega_1, \Omega_2) \subset \mathcal{T}_0(\Omega_1, \Omega_2) \subset \mathcal{T}(\Omega_1, \Omega_2) \tag{3.4}$$

Furthermore, $\mathcal{T}_r(\Omega_1, \Omega_2)$ *and* $\mathcal{T}_0(\Omega_1, \Omega_2)$ *are dense in* $\mathcal{T}(\Omega_1, \Omega_2)$ *(with respect to the topology of pointwise convergence).*

Proof: Equation (3.4) is obvious from the definitions. The second statement is a special case of (Le Cam, 1986, Theorem 1.4.1):

In (Le Cam, 1986, Theorem 1.4.1) put $L = \mathrm{ba}(\Omega_1, \mathcal{A}_1)$, $D = \Omega_2$, $\Gamma = \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$ and $H = M$. The transitions which are finitely supported and (Γ, H) continuous are dense in $\mathcal{T}(\Omega_1, \Omega_2)$ with respect to the topology of uniform convergence on the elements of $\mathcal{K}$ (as defined in (Le Cam, 1986, p. 7)).

Since $\{\mu_1\} \times \{f_2\} \in \mathcal{K}$ for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and every $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$, the topology of pointwise convergence is weaker than the topology of uniform convergence on the elements of $\mathcal{K}$.

Hence, the transitions which are finitely supported and (Γ, H) continuous are also dense in $\mathcal{T}(\Omega_1, \Omega_2)$ with respect to the topology of pointwise convergence and it suffices to prove the following statement: Every finitely supported and (Γ, H) continuous transition (according to (Le Cam, 1986, p. 6f)) is a restricted randomization (according to Definition 3.7).

The latter statement is shown by Proposition 3.37 below. □

Theorem 3.9 and Theorem 3.10 are due to L. Le Cam. Results from Le Cam (1986) can be used in this book because the setup here is a special case of the general setup in Le Cam (1986). However, for the reader who is not familiar with the general setup, it is very hard (or even impossible) to look up these results in Le Cam (1986). Therefore, the connection of the setup in this book and the general setup in Le Cam (1986) and Le Cam (1964) is explained in Section 3.4.

The present subsection ends with a convenient characterization of ordinary randomizations.

Proposition 3.11 *Let Ω_1 be a set with algebra $\mathcal{A}_1$ and let Ω_2 be another set with algebra $\mathcal{A}_2$. Let $\sigma \in \mathcal{T}(\Omega_1, \Omega_2)$ be a generalized randomization. Then, the following statements are all equivalent:*

- ***a)*** *σ is an ordinary randomization.*
- ***b)*** *There is a map $T : \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \rightarrow \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$ which is linear, positive ($T(f_2) \geq 0 \ \forall f_2 \geq 0$) and normalized ($T(I_{\Omega_2}) = I_{\Omega_1}$) such that σ is the adjoint operator of T.*
- ***c)*** *σ is continuous with respect to the $\mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$ - topology on* $\mathrm{ba}(\Omega_1, \mathcal{A}_1)$ *and the $\mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$ - topology on* $\mathrm{ba}(\Omega_2, \mathcal{A}_2)$.

Proof:

(a) ⇒ (b): As already stated above, this is a direct consequence of the definition. Put $T(f_2)(\omega_1) = \tau_{\omega_1}[f_2]$ where τ is a finitely additive Markov kernel which defines σ via (3.2).

(a) ⇐ (b): T defines a finitely additive Markov kernel via $\tau_{\omega_1}[I_{A_2}] = T(I_{A_2})(\omega_1)$. Then,

$$\sigma(\mu_1)[I_{A_2}] = \mu_1\big[T(I_{A_2})\big] = \int\!\!\int I_{A_2}(\omega_2)\,\tau_{\omega_1}(d\omega_2)\,\mu_1(d\omega_1)$$

for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and $A_2 \in \mathcal{A}_2$. According to the definition of $\mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$, this implies

$$\sigma(\mu_1)[f_2] \;=\; \int\int f_2(\omega_2)\,\tau_{\omega_1}(d\omega_2)\,\mu_1(d\omega_2)$$

for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$.

(b) ⇒ (c): All topological terms within this proof are with respect to the topologies mentioned in Proposition 3.11.
Let $(\mu_{1,\gamma})_{\gamma\in D}$ be a net in $\mathrm{ba}(\Omega_1, \mathcal{A}_1)$ which converges to some $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$. This implies that, for every $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$,

$$\sigma(\mu_{1,\gamma})[f_2] \;=\; \mu_{1,\gamma}\big[T(f_2)\big] \;\underset{\gamma}{\longrightarrow}\; \mu_1\big[T(f_2)\big] \;=\; \sigma(\mu_1)[f_2]$$

That is, $\sigma(\mu_{1,\gamma}) \underset{\gamma}{\longrightarrow} \sigma(\mu_1)$ according to Theorem 8.24 b).

(b) ⇐ (c): According to (Dunford and Schwartz, 1958, Exercise VI.9.13), there is a norm-continuous linear functional $T : \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \rightarrow \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$ such that σ is the adjoint operator of T.
T is positive because $T(f_2)(\omega_1) = \delta_{\omega_1}\big[T(f_2)\big] = \sigma(\delta_{\omega_1})[f_2] \geq 0$ for every $f_2 \geq 0$ where δ_{ω_1} denotes the Dirac measure.
T is normalized because $T(I_{\Omega_2})(\omega_1) = \delta_{\omega_1}\big[T(I_{\Omega_2})\big] = \sigma(\delta_{\omega_1})[I_{\Omega_2}] = 1$. □

3.3.1.2 Generalized decision procedures

As explained in Section 3.2, decision procedures (called randomized decision functions) are defined via Markov kernels. In the previous subsection, generalized randomizations were defined as generalizations of Markov kernels. So, it is suggesting to use this definition in order to generalize randomized decision functions:

Definition 3.12 *Let* $(\mathbb{D},\mathcal{D})$ *be a decision space and* $(\mathcal{Y},\mathcal{B})$ *a sample space. A* (generalized) decision procedure *is a generalized randomization*

$$\sigma \ : \ \mathrm{ba}(\mathcal{Y},\mathcal{B}) \ \rightarrow \ \mathrm{ba}(\mathbb{D},\mathcal{D})$$

In order to define the risk function of such a generalized decision procedure

$$\sigma \ : \ \mathrm{ba}(\mathcal{Y},\mathcal{B}) \ \rightarrow \ \mathrm{ba}(\mathbb{D},\mathcal{D})$$

let

$$(W_\theta)_{\theta\in\Theta} \ \subset \ \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$$

be a loss function and $(Q_\theta)_{\in\theta\in\Theta}$ be a precise model on the sample space $(\mathcal{Y},\mathcal{B})$. Then, the risk function of σ is defined to be

$$\Theta \ \rightarrow \ \mathbb{R}\,, \qquad \theta \ \mapsto \ \sigma(P_\theta)[W_\theta]$$

Accordingly, the risk function of σ for an imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$ is defined to be

$$\Theta \ \rightarrow \ \mathbb{R}\,, \qquad \theta \ \mapsto \ \sup_{P_\theta\in\mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]$$

where $\mathcal{M}_\theta$ is the credal set which corresponds to Q_θ for every $\theta\in\theta$.

Of course, these definitions reduce to the usual ones if the decision procedure σ is defined by an ordinary randomization; confer also Section 3.2.

In order to unify terminology, the following definitions are used, too:

Definition 3.13 *Let* $(\mathbb{D},\mathcal{D})$ *be a decision space and* $(\mathcal{Y},\mathcal{B})$ *a sample space. A* restricted / ordinary decision procedure *is a restricted / ordinary randomization*

$$\sigma \ : \ \mathrm{ba}(\mathcal{Y},\mathcal{B}) \ \rightarrow \ \mathrm{ba}(\mathbb{D},\mathcal{D})$$

That is, every ordinary decision procedure corresponds to a randomized decision function and vice versa.

3.3.2 Sufficiency and equivalence of imprecise models

3.3.2.1 Definitions and basic properties

Sufficiency is not only a very important concept in statistics but also in decision theory. The following definition is an analog to the corresponding definition in Buja (1984).

Definition 3.14 (Sufficient) *Let* Θ *be any index set. Let* $\mathcal{Y}$ *be a set with algebra* $\mathcal{B}$ *and let* $(Q_\theta)_{\theta\in\Theta}$ *be a precise model on* $(\mathcal{Y},\mathcal{B})$*. Let* $\mathcal{X}$ *be another set with algebra* $\mathcal{A}$ *and let* $(P_\theta)_{\theta\in\Theta}$ *be a precise model on* $(\mathcal{X},\mathcal{A})$ *.*
$(P_\theta)_{\theta\in\Theta}$ *is called* sufficient for $(Q_\theta)_{\theta\in\Theta}$ *if there is a generalized randomization* $\sigma\in\mathcal{T}(\mathcal{X},\mathcal{Y})$ *so that* $\sigma(P_\theta)=Q_\theta \ \ \forall\,\theta\in\Theta$ *.*

This definition of “sufficiency” essentially goes back to Blackwell (1951). It does not strictly coincide with the more common definition in terms of conditional expectations but, under suitable assumptions of regularity, the definitions do coincide (cf. Heyer (1973)). At least, if the randomization σ is an ordinary randomization, which is defined by a randomized function $x \mapsto \tau_x$, it has a very descriptive interpretation:
Let x be an observation distributed according to P_θ. After observing x, start an auxiliary random experiment according to τ_x. Then, the outcome y of the auxiliary random experiment is distributed according to Q_θ. That is, if we have observations of the model $(P_\theta)_{\theta\in\Theta}$, we can artificially generate observations of the model $(Q_\theta)_{\theta\in\Theta}$ “by coin tossing”. This illustrates that (at least in case of an ordinary randomization) the model $(Q_\theta)_{\theta\in\Theta}$ cannot be more informative than the sufficient model $(P_\theta)_{\theta\in\Theta}$.

Using this definition of sufficiency, we can define an equivalence relation on the set of all precise models

$$\Big\{\big(\mathcal{Z},\mathcal{C},(P_\theta)_{\theta\in\Theta}\big)\ \Big|\ \mathcal{Z} \text{ a set with algebra } \mathcal{C}\,,\ \ (P_\theta)_{\theta\in\Theta} \subset \mathrm{ba}_1^+(\mathcal{Z},\mathcal{C})\Big\}$$

Note that, while the sample space $(\mathcal{Z},\mathcal{C})$ may change, the index set Θ is fixed.

Definition 3.15 (Equivalence of models) *Let Θ be any index set. Let $\mathcal{Y}$ be a set with algebra $\mathcal{B}$ and let $(Q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y},\mathcal{B})$. Let $\mathcal{X}$ be another set with algebra $\mathcal{A}$ and let $(P_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{X},\mathcal{A})$.*
$(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ are called equivalent *if they are mutually sufficient.*

That is, $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ are equivalent if and only if there are some $\sigma \in \mathcal{T}(\mathcal{X},\mathcal{Y})$, $\rho \in \mathcal{T}(\mathcal{Y},\mathcal{X})$ so that $\sigma(P_\theta) = Q_\theta\ \forall\,\theta\in\Theta$ and $\rho(Q_\theta) = P_\theta\ \forall\,\theta\in\Theta$. This definition of equivalence is in accordance with Le Cam’s definition (cf. Proposition 3.39). The descriptive interpretation of sufficiency already indicates that equivalent models essentially coincide from a decision theoretic point of view.

Now, we turn over to imprecise models. Recall from Notation 3.2, that the term

$$\text{“}\big(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta}\big)\ \text{ is an imprecise model”}$$

means that $\mathcal{Y}$ is a set with algebra $\mathcal{B}$ and $(\overline{Q}_\theta)_{\theta\in\Theta}$ is an imprecise model on $(\mathcal{Y},\mathcal{B})$. An analogous notation is used for precise models.
The following definition is again an analog to the corresponding definition of worst-case-sufficiency in Buja (1984).

Definition 3.16 (Worst-case-sufficient) *Let Θ be any index set. Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model with corresponding family of credal sets $(\mathcal{M}_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$. Let $(\mathcal{X},\mathcal{A},(P_\theta)_{\theta\in\Theta})$ be a precise model.*
$(P_\theta)_{\theta\in\Theta}$ is called worst-case-sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$ *if there is precise model $(Q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ such that $(P_\theta)_{\theta\in\Theta}$ is sufficient for $(Q_\theta)_{\theta\in\Theta}$.*

That is, $(P_\theta)_{\theta\in\Theta}$ is worst-case-sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$ if and only if there is some $\sigma \in \mathcal{T}(\mathcal{X},\mathcal{Y})$ so that $\forall\,\theta\in\Theta$

$$\sigma(P_\theta)[g] \ \le\ \overline{Q}_\theta[g]\,, \qquad \forall\, g \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$$

Indeed, worst-case-sufficiency is a very weak form of sufficiency but it is often enough because we have chosen a worst case consideration in decision theory under imprecise probabilities; cf. Section 3.1.

Finally, sufficiency can also be defined for imprecise models:

Definition 3.17 (Sufficient) *Let Θ be any index set. Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model with corresponding family of credal sets $(\mathcal{M}_\theta)_{\theta\in\Theta}$. Let $(\mathcal{X},\mathcal{A},(\overline{P}_\theta)_{\theta\in\Theta})$ be an imprecise model with corresponding family of credal sets $(\mathcal{N}_\theta)_{\theta\in\Theta}$.*
$(\overline{P}_\theta)_{\theta\in\Theta}$ is called sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$ *if there is a generalized randomization $\sigma\in\mathcal{T}(\mathcal{X},\mathcal{Y})$ such that $\forall\,\theta\in\Theta$*

$$\sup_{P_\theta\in\mathcal{N}_\theta}\sigma(P_\theta)[g] \;=\; \overline{Q}_\theta[g]\,, \qquad \forall\,g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \tag{3.5}$$

Proposition 3.18 *Let Θ be any index set. Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model with corresponding family of credal sets $(\mathcal{M}_\theta)_{\theta\in\Theta}$. Let $(\mathcal{X},\mathcal{A},(\overline{P}_\theta)_{\theta\in\Theta})$ be an imprecise model with corresponding family of credal sets $(\mathcal{N}_\theta)_{\theta\in\Theta}$.*
Then, $(\overline{P}_\theta)_{\theta\in\Theta}$ is sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$ if and only if there is a generalized randomization $\sigma\in\mathcal{T}(\mathcal{X},\mathcal{Y})$ such that

$$c\ell\Big(\sigma(\mathcal{N}_\theta)\Big) \;=\; \mathcal{M}_\theta \qquad \forall\,\theta\in\Theta$$

where $c\ell$ denotes the closure with respect to the $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ - topology of $\mathrm{ba}(\mathcal{Y},\mathcal{B})$.

Proof: Proposition 2.15 and (3.5) imply that the credal set $\mathcal{M}_\theta$ of $\overline{Q}_\theta$ is equal to $c\ell\,\mathrm{co}\big(\sigma(\mathcal{N}_\theta)\big)$. Convexity of $\mathcal{N}_\theta$ and linearity of σ imply that $\sigma(\mathcal{N}_\theta)$ is convex. Hence, $c\ell\big(\sigma(\mathcal{N}_\theta)\big)$ is the convex closure of $\sigma(\mathcal{N}_\theta)$ according to (Dunford and Schwartz, 1958, Theorem V.2.1). That is, $c\ell\big(\sigma(\mathcal{N}_\theta)\big)=c\ell\,\mathrm{co}\big(\sigma(\mathcal{N}_\theta)\big)=\mathcal{M}_\theta$.
The converse statement is trivial. □

Remark 3.19 *If the generalized randomization σ in Definition 3.17 is even an ordinary randomization, then*

$$\sigma(\mathcal{N}_\theta) \;=\; \mathcal{M}_\theta$$

is the credal set of $\overline{Q}_\theta$ for every $\theta\in\Theta$. This follows from Proposition 3.18 and Proposition 3.11.

Analogously to precise models, we can also define equivalence for imprecise models:

Definition 3.20 *Let Θ be any index set. Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ and $(\mathcal{X},\mathcal{A},(\overline{P}_\theta)_{\theta\in\Theta})$ be imprecise models.*
$(\overline{P}_\theta)_{\theta\in\Theta}$ and $(\overline{Q}_\theta)_{\theta\in\Theta}$ are called equivalent *if they are mutually sufficient.*

Remark 3.21 *Since probability charges are special cases of coherent upper previsions, precise models $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ are special cases of imprecise models. Hence, the terms "sufficient" and "equivalent" have been defined twice for $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ – considered as precise models in Definition 3.14 and Definition 3.15, considered as imprecise models in Definition 3.17 and Definition 3.20. However, it is obvious that the respective definitions coincide in this case.*

Every imprecise model is equivalent to a uniquely defined imprecise model on a compact Hausdorff space which consists of upper expectations – namely the canonical Stone representation:

Theorem 3.22 (Equivalence of the canonical Stone representation)
Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model. Let $(\Xi,\mathfrak{B}_0)$ be the canonical Stone space of $(\mathcal{Y},\mathcal{B})$ and, for each $\theta\in\Theta$, let $\overline{S}_{0,\theta}$ be the canonical Stone representation of $\overline{Q}_\theta$ (according to Theorem 2.44).
Then, $(\overline{Q}_\theta)_{\theta\in\Theta}$ and $(\overline{S}_{0,\theta})_{\theta\in\Theta}$ are equivalent.

Proof: For $\theta \in \Theta$, let $\mathcal{M}_\theta$ be the credal sets of $\overline{Q}_\theta$ and $\mathcal{N}_{0,\theta}$ the credal set of $\overline{S}_{0,\theta}$.

According to Theorem 2.40, there is an L-space isomorphism

$$\phi_0 \;:\; \mathrm{ba}(\mathcal{Y},\mathcal{B}) \;\rightarrow\; \mathrm{ca}(\Xi,\mathfrak{B}_0)$$

such that

$$\sup_{Q_\theta \in \mathcal{M}_\theta} \phi_0(Q_\theta)[h_0] \;=\; \overline{S}_{0,\theta}[h_0] \qquad \forall\, h_0 \in \mathcal{L}_\infty(\Xi,\mathfrak{B}_0)$$

for every $\theta \in \Theta$. It is easy to see that an L-space isomorphism is always a transition[3]. Hence, ϕ_0 is indeed a transition – i.e. a generalized randomization – and $(\overline{Q}_\theta)_{\theta\in\Theta}$ is sufficient for $(\overline{S}_{0,\theta})_{\theta\in\Theta}$.

Conversely, put

$$\sigma(\nu_0)[f] \;=\; \nu_0\big[\xi(f)\big] \qquad \forall\, \nu_0 \in \mathrm{ba}(\Xi,\mathfrak{B}_0)\,, \quad \forall\, f \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$$

where ξ is the canonical Stone kernel. This defines a generalized randomization

$$\sigma \;:\; \mathrm{ba}(\Xi,\mathfrak{B}_0) \;\rightarrow\; \mathrm{ba}(\mathcal{Y},\mathcal{B})\,, \qquad \nu_0 \;\mapsto\; \sigma(\nu_0)$$

and, according to Theorem 2.44,

$$\overline{Q}_\theta[f] \overset{(2.39)}{=} \overline{S}_{0,\theta}\big[\xi(f)\big] \;=\; \sup_{S_{0,\theta}\in\mathcal{N}_{0,\theta}} S_{0,\theta}\big[\xi(f)\big] \;=\; \sup_{S_{0,\theta}\in\mathcal{N}_{0,\theta}} \sigma(\nu_0)[f]$$

for every $f \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$. Hence, $(\overline{S}_{0,\theta})_{\theta\in\Theta}$ is sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$. □

For the above theorem, it is crucial that the canonical Stone space $(\Xi,\mathfrak{B}_0)$ for a coherent upper prevision $\overline{Q}_\theta$ only depends on $(\mathcal{Y},\mathcal{B})$ and does not depend on $\overline{Q}_\theta$. Therefore, $(\Xi,\mathfrak{B}_0)$ does not depend on θ and we are able to define an imprecise model $(S_{0,\theta})_{\theta\in\Theta}$ on $(\Xi,\mathfrak{B}_0)$.

As seen in case of the canonical Stone representation, this notion of sufficiency is an interesting theoretical tool. In addition, Subsection 3.3.2.2 contains a nice example for sufficiency of imprecise models which shows that this concept can also be applied in real statistical situations.

In standard mathematical statistics, sufficiency is usually defined by conditional expectations. The concept of conditional expectations which arise in this definition is a rather abstract measure theoretic concept which deeply relies on σ-additivity. Since the definition of conditional expectations for imprecise previsions is a complicated matter of its own which rises many problems even in case of finite sample spaces[4], one might think that a definition of sufficiency for imprecise previsions is far of the scope of present research. However, as seen above, a generalization of the definitions in Blackwell (1951) and Buja (1984) really leads to a general definition of sufficiency for imprecise previsions which avoids the problems connected with conditional expectations.

[3] This is a direct consequence of the definitions and (2.10).

[4] The present state of the art is that there seems to be no definition of conditional expectations for imprecise previsions which is satisfactory in every situation. Instead, different situations ask for different definitions of conditional expectations; cf. e.g. Weichselberger and Augustin (2003).

3.3.2.2 Examples

a) Classical sufficiency

Let $\mathcal{Y}$ be a Polish space with Borel-σ-algebra $\mathfrak{B}$, and let $(Q_\theta)_{\theta\in\Theta} \subset \mathrm{ca}_1^+(\mathcal{Y},\mathfrak{B})$ be a precise model which consists of probability measures Q_θ on $(\mathcal{Y},\mathfrak{B})$. Θ may be any index set.
Furthermore, let $\mathcal{X}$ be a set with σ-algebra $\mathcal{A}$, and let

$$X \;:\; \mathcal{Y} \;\to\; \mathcal{X}$$

be a $\mathfrak{B}/\mathcal{A}$ - measurable map.
That is, we are faced with the following situation:

$$\big(\mathcal{Y},\mathfrak{B},(Q_\theta)_{\theta\in\Theta}\big) \quad \xrightarrow{\;\;X\;\;} \quad \big(\mathcal{X},\mathcal{A},(Q_\theta^X)_{\theta\in\Theta}\big)$$

where Q_θ^X denotes the image measure

$$Q_\theta^X \;:\; A \;\mapsto\; Q_\theta\big(X^{-1}(A)\big)$$

on $(\mathcal{X},\mathcal{A})$; cf. e.g. (Hoffmann-Jørgensen, 1994a, § 1.44).

Assume that X is sufficient in the usual sense of mathematical statistics; cf. e.g. (Shao, 2003, Definition 2.4). That is, the conditional distribution given $X = x$ with respect to Q_θ

$$\mathfrak{B} \;\to\; [0,1]\,, \qquad B \;\mapsto\; Q_\theta(B|X=x)$$

does not depend on θ. Since $(\mathcal{Y},\mathfrak{B})$ is assumed to be a Polish space, there is a regular version of the conditional expectation.[5] That is, there is a Markov kernel

$$\tau \;:\; \mathfrak{B}\times\mathcal{X} \;\to\; \mathbb{R}\,, \qquad (B,x) \;\mapsto\; \tau_x(B)$$

such that

$$\tau_x(B) \;=\; Q_\theta(B|X=x) \qquad \text{for } Q_\theta^X \text{ - almost every } x\in\mathcal{X}\,, \qquad \forall\,\theta\in\Theta$$

Put

$$P_\theta \;:=\; Q_\theta^X\,, \qquad \forall\,\theta\in\Theta$$

Then, $\big(\mathcal{X},\mathcal{A},(P_\theta)_{\theta\in\Theta}\big)$ is sufficient for $\big(\mathcal{Y},\mathfrak{B},(Q_\theta)_{\theta\in\Theta}\big)$ according to Definition 3.14 and 3.17 because the Markov kernel τ defines an ordinary randomization σ such that

$$\begin{aligned}\sigma(P_\theta)[g] \;&=\; \int_{\mathcal{X}}\int_{\mathcal{Y}} g(y)\,\tau_x(dy)P_\theta(dx) \;=\; \int_{\mathcal{X}}\int_{\mathcal{Y}} g(y)\,Q_\theta(dy|X\!=\!x)\,Q_\theta^X(dx)\\ &=\; \int_{\mathcal{Y}} g(y)Q_\theta(dy) \;=\; Q_\theta[g] \qquad \forall\,g\in\mathcal{L}_\infty(\mathcal{Y},\mathfrak{B})\end{aligned}$$

for every $\theta\in\Theta$.

b) Robust statistics

[5] Confer (Bauer, 1996, Theorem 44.3), for example.

Let $\big(\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Theta}\big)$ be a precise model and $\mathcal{U}(Q_\theta)$ be a neighborhood of Q_θ as common in robust statistics; cf. e.g. (Rieder, 1994, § 4.2.1). Put

$$\overline{Q}_\theta[g] \;=\; \sup_{K_\theta \in \mathcal{U}(Q_\theta)} K_\theta[g] \qquad \forall\, g \in \mathcal{L}_\infty(\mathcal{Y}, \mathfrak{B})\,, \qquad \theta \in \Theta$$

If $\big(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}\big)$ is a precise model which is sufficient for $\big(\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Theta}\big)$, then it is also worst-case-sufficient for $\big(\mathcal{Y}, \mathfrak{B}, (\overline{Q}_\theta)_{\theta\in\Theta}\big)$.

c) Parametrically generated coherent upper prevision

Similar to parametrically generated F-probabilities in (Weichselberger, 2001, p. 131ff), parametrically generated coherent upper expectations may be defined:

Definition 3.23 (Parametrically generated coherent upper previsions)
Let $\mathcal{Y}$ be a set with algebra $\mathcal{B}$. Let $(Q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y}, \mathcal{B})$ where Θ is any index set. Furthermore, let $H \subset \Theta$ be any subset of the index set.
A coherent upper prevision $\overline{Q}_H$ is called parametrically generated by H (with respect to $(Q_\theta)_{\theta\in\Theta}$) *if it is given by*

$$\overline{Q}_H[g] \;=\; \sup_{\theta\in H} Q_\theta[g]\,, \qquad g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$$

Usually, we have $\Theta \subset \mathbb{R}^k$ and

$$H \;=\; [\underline{\theta}_1, \overline{\theta}_1] \times [\underline{\theta}_2, \overline{\theta}_2] \times \cdots \times [\underline{\theta}_k, \overline{\theta}_k] \;\subset\; \Theta$$

for some suitable real numbers $\underline{\theta}_j \;\le\; \overline{\theta}_j$, $j = 1, \dots, k$.

An imprecise model which consists of such parametrically generated coherent upper previsions is called *parametrically generated imprecise model*:

Definition 3.24 (Parametrically generated imprecise model)
Let $\mathcal{Y}$ be a set with algebra $\mathcal{B}$. Let $(Q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y}, \mathcal{B})$ where Θ is any index set. Furthermore, let $\mathcal{H} \subset 2^\Theta$ be any subset of the power set of Θ.
An imprecise model $(\overline{Q}_H)_{H\in\mathcal{H}}$ is called parametrically generated by $\mathcal{H}$ (with respect to $(Q_\theta)_{\theta\in\Theta}$ *if the coherent upper prevision $\overline{Q}_H$ is parametrically generated by H with respect to $(Q_\theta)_{\theta\in\Theta}$ for every $H \in \mathcal{H}$.*

Now, consider the situation in part a) again:

$\mathcal{Y}$ is a Polish space with Borel-σ-algebra $\mathfrak{B}$, $(Q_\theta)_{\theta\in\Theta} \;\subset\; \mathrm{ca}_1^+(\mathcal{Y}, \mathfrak{B})$ is a precise model which consists of probability measures Q_θ on $(\mathcal{Y}, \mathfrak{B})$.
$\mathcal{X}$ is a set with σ-algebra $\mathcal{A}$, and $X:\; \mathcal{Y} \to \mathcal{X}$ is a $\mathcal{A}/\mathfrak{B}$ - measurable map. That is:

$$\big(\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Theta}\big) \;\xrightarrow{\;X\;}\; \big(\mathcal{X}, \mathcal{A}, (Q_\theta^X)_{\theta\in\Theta}\big)$$

Assume that X is sufficient in the usual sense of mathematical statistics; cf. e.g. (Shao, 2003, Definition 2.4). Then, it was shown in part a) that $\big(\mathcal{X}, \mathcal{A}, (Q_\theta^X)_{\theta\in\theta}\big)$ is sufficient for $\big(\mathcal{Y}, \mathfrak{B}, (Q_{\theta\in\theta}\big)$ in the sense of Definition 3.17. The following theorem states that this sufficiency also implies sufficiency of the respective parametrically generated imprecise models:

Theorem 3.25 *In the above setting, let X be sufficient in the sense of (Shao, 2003, Definition 2.4) and $\mathcal{H} \subset 2^{\Theta}$ be any subset of the power set of Θ. Furthermore, let $\big(\mathcal{X}, \mathcal{A}, (\overline{P}_H)_{H\in\mathcal{H}}\big)$ be the imprecise model which is parametrically generated by $\mathcal{H}$ with respect to $(Q_\theta^X)_{\theta\in\theta}$ and let $\big(\mathcal{Y}, \mathfrak{B}, (\overline{Q}_H)_{H\in\mathcal{H}}\big)$ be the imprecise model which is parametrically generated by $\mathcal{H}$ with respect to $(Q_\theta)_{\theta\in\theta}$.*
Then, $(\overline{P}_H)_{H\in\mathcal{H}}$ is sufficient for $(\overline{Q}_H)_{H\in\mathcal{H}}$.

Proof: According to part a) of the present subsection, there is an ordinary randomization

$$\sigma \;\in\; \mathcal{T}_0(\mathcal{X}, \mathcal{Y})$$

such that

$$\sigma\big(Q_\theta^X\big) \;=\; Q_\theta \qquad \forall\, \theta \in \Theta$$

For $H \in \mathcal{H}$, let $\mathcal{N}_H$ be the credal set of $\overline{P}_H$ on $(\mathcal{X}, \mathcal{A})$. Endow $\mathrm{ba}(\mathcal{Y}, \mathfrak{B})$ with the $\mathcal{L}_\infty(\mathcal{Y}, \mathfrak{B})$ - topology and $\mathrm{ba}(\mathcal{X}, \mathcal{A})$ with the $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ - topology. Then,

$$\mathrm{ba}(\mathcal{X}, \mathcal{A}) \;\to\; \mathbb{R}\,, \qquad \mu \;\mapsto\; \sigma(\mu)[g]$$

is linear and continuous for every $g \in \mathcal{L}_\infty(\mathcal{Y}, \mathfrak{B})$ according to Proposition 3.11 and Lemma 8.25. This implies

$$\overline{Q}_H[g] \;=\; \sup_{\theta\in H} Q_\theta[g] \;=\; \sup_{\theta\in H} \sigma\big(Q_\theta^X\big)[g] \;=\; \sup_{P_H\in\mathcal{N}_H} \sigma(P_H)[g]$$

where the last equality follows from Proposition 2.15 and Lemma 8.29. □

Remark 3.26 *The definition of sufficiency of coherent upper previsions can easily be rewritten into an analogous definition in case of upper expectations. Then, an analog of Theorem 3.25 can be proven for upper expectations following the lines of the above proof where Proposition 2.15 is replaced by Proposition 2.23.*

In order to get a better impression of the above theorem, we may consider a more concrete example:

In an ideal situation, the outcomes of an experiment may be distributed according to a one-dimensional normal distribution

$$y_1,\; y_2,\; y_3,\; \ldots\,, y_{100} \quad\sim\quad \mathcal{N}(\theta_1, \theta_2)$$

Here, we have 100 independently, identically distributed observations $y_i \in \mathbb{R}$ and $\theta = (\theta_1, \theta_2)$ is the parameter where $\theta_1 \in (-\infty, \infty)$ denotes the mean and $\theta_2 \in (0, \infty)$ denotes the variance of the normal distribution. The assumptions imply that the observation is

$$y \;:=\; (y_1, y_2, \ldots, y_{100}) \;\sim\; \mathcal{N}(\theta_1, \theta_2)^{\otimes 100} \;=\; \mathcal{N}_{100}\big(a(\theta_1), B(\theta_2)\big)$$

where

$$a(\theta_1) \;=\; \begin{pmatrix} \theta_1 \\ \theta_1 \\ \cdot \\ \cdot \\ \cdot \\ \theta_1 \end{pmatrix} \in \mathbb{R}^{100}\,, \quad b(\theta_2) \;=\; \begin{pmatrix} \theta_2 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \theta_2 & 0 & \cdot & \cdot & 0 \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ \cdot & & & & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & \theta_2 \end{pmatrix} \in \mathbb{R}^{100\times 100}$$

So, the parametric precise model is $\big(\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Theta}\big)$ where

$$\mathcal{Y} \;=\; \mathbb{R}^{100}\,, \qquad Q_\theta \;=\; \mathcal{N}_{100}\big(a(\theta_1), b(\theta_2)\big)\,, \qquad \Theta \;=\; (-\infty,\infty)\times(0,\infty)$$

In a less idealized situation, only a *parametrically generated* imprecise model is assumed: Let $\mathcal{H}$ be a subset of 2^Θ such that each $H\in\mathcal{H}$ is of form

$$H \;=\; [\underline{\theta}_1, \overline{\theta}_1] \times [\underline{\theta}_2, \overline{\theta}_2] \;\subset\; \Theta$$

for some suitable real numbers $\underline{\theta}_j \;\le\; \overline{\theta}_j \;,\; j\in\{1,2\}\,.$
Then, the parametrically generated imprecise model is given by

$$\overline{Q}_H[g] \;=\; \sup_{\theta\in H} \int_{\mathbb{R}^{100}} g\, d\mathcal{N}_{100}\big(a(\theta_1), b(\theta_2)\big)$$

No matter if we consider the precise model $\big(\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Theta}\big)$ or the imprecise model $\big(\mathcal{Y}, \mathfrak{B}, (\overline{Q}_H)_{H\in\mathcal{H}}\big)$, the sample space is

$$\mathcal{Y} \;=\; \mathbb{R}^{100}$$

which is quite intractable. However, the precise model becomes tractable by the well-known fact that the map

$$X \;:\; \mathbb{R}^{100} \;\longrightarrow\; \mathbb{R}^2\,, \qquad (y_1, y_2, \ldots, y_{100}) \;\mapsto\; \left(\sum_{i=1}^{100} y_i \;,\; \sum_{i=1}^{100} y_i^2\right)$$

is sufficient. That is, we do not need the quite large sample space $\mathcal{Y} \;=\; \mathbb{R}^{100}$; it is enough to consider the precise model

$$P_\theta \;=\; Q_\theta^X\,, \qquad \theta\in\Theta$$

on the sample space

$$\mathcal{X} \;=\; (-\infty,\infty)\times(0,\infty) \;\subset\; \mathbb{R}^2$$

Now, Theorem 3.25 states that the same is true also in case of the parametrically generated imprecise model: It is enough to consider the parametrically generated imprecise model

$$\overline{P}_H[f] \;=\; \sup_{\theta\in H} P_\theta[f]\,, \qquad \forall\, f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})\,, \qquad H\in\mathcal{H}$$

on the smaller sample space

$$\mathcal{X} \;=\; (-\infty,\infty)\times(0,\infty) \;\subset\; \mathbb{R}^2$$

Another example where the above introduced notion of sufficiency for imprecise models can be used is the so-called “Imprecise Dirichlet Model” which attracts an amazing amount of attention in the theory of imprecise probabilities. The Imprecise Dirichlet Model is an imprecise model which is parametrically generated by the Dirichlet distribution. Since the precise Dirichlet model is an exponential family, valuable statistics are at hand which are sufficient in the usual sence of mathematical statistics. According to Theorem 3.25, these statistics are also sufficient for the Imprecise Dirichlet Model.

3.3.3 Standard Models

In Subsection 3.3.2, we have defined an equivalence relation on the precise models with a fixed index set Θ. Each equivalence class contains a uniquely defined representative (called standard model later on) which has some nice properties. As stated in Subsection 3.3.2, equivalent models coincide from a decision theoretic point of view. Therefore, every decision problem coincides with a "standard decision problem" where a standard model is involved; properties of the original decision problem can be deduced from the corresponding "standard decision problem"; confer Section 4.2.

Let the index set Θ be finite with cardinality n now. Furthermore, let $\mathcal{X}$ be a set with σ-algebra $\mathcal{A}$ and $(P_\theta)_{\theta\in\Theta}$ a precise model where each P_θ is not only a probability charge but even a probability *measure*. In this situation, the standard model can be defined in the following way:

Put

$$P \;=\; \frac{1}{n}\sum_{\theta\in\Theta} P_\theta \;\in\; \mathrm{ca}_1^+(\mathcal{X},\mathcal{A})$$

Hence, for each P_θ, there is a P-density β_θ such that

$$P_\theta(A) \;=\; \int_A \beta_\theta\, dP \qquad \forall\, A\in\mathcal{A}$$

The maps β_θ can be chosen such that $\beta_\theta \geq 0$ and $\displaystyle\frac{1}{n}\sum_{\theta\in\Theta}\beta_\theta \equiv 1$ because

$$\int_A \sum_{\theta\in\Theta}\beta_\theta\, dP \;=\; \sum_{\theta\in\Theta}\int_A \beta_\theta\, dP \;=\; \sum_{\theta\in\Theta} P_\theta(A) \;=\; n\cdot P(A) \;=\; n\int_A 1\, dP$$

for every $A\in\mathcal{A}$. Therefore, it can be assumed without loss of generality that

$$\beta(x) \;:=\; \frac{1}{n}\big(\beta_{\theta_1}(x)\,,\,\dots\,,\,\beta_{\theta_n}(x)\big) \;\in\; \mathcal{U} \qquad \forall\, x\in\mathcal{X}$$

where

$$\mathcal{U} \;:=\; \big\{u\in\mathbb{R}^n \;\big|\; u=(u_{\theta_1},\dots,u_{\theta_n})\,,\;\; u_\theta\geq 0\;\;\forall\,\theta\in\Theta\,,\;\; u_{\theta_1}+\dots+u_{\theta_n}=1\big\}$$

Put $\mathcal{C} := \mathfrak{B}^{\otimes n}\cap\mathcal{U}$ where $\mathfrak{B}^{\otimes n}$ is the Borel-σ-algebra of $\mathbb{R}^n$. Then,

$$\beta \;:\; \mathcal{X} \;\to\; \mathcal{U}\,, \qquad x \;\mapsto\; \beta(x)$$

is an $\mathcal{A}/\mathcal{C}$ - measurable function and we can define the image measures

$$S \;:=\; \beta(P) \quad \text{ where } \quad S(C)=\beta(P)(C)=P\big(\beta^{-1}(C)\big) \quad \forall\, C\in\mathcal{C} \tag{3.6}$$

and, for each $\theta\in\Theta$,

$$S_\theta \;:=\; \beta(P_\theta) \quad \text{ where } \quad S_\theta(C)=\beta(P_\theta)(C)=P_\theta\big(\beta^{-1}(C)\big) \quad \forall\, C\in\mathcal{C} \tag{3.7}$$

Of course, these image measures are probability measures on $(\mathcal{U},\mathcal{C})$. Let $\iota_\theta : \mathcal{U}\to[0,1]$, $u\mapsto u_\theta$ denote the projection of u on the θ-component u_θ. Then, $\beta_\theta = n\cdot(\iota_\theta\circ\beta)$ and, therefore, the definitions imply

$$S_\theta(C) \;=\; \int_C n\iota_\theta\, dS \qquad \forall\, C\in\mathcal{C} \tag{3.8}$$

Since $\mu \mapsto \beta(\mu)$ [6] is a randomization and $\beta(P_\theta) = S_\theta$ for every $\theta \in \Theta$, the precise model $(P_\theta)_{\theta\in\Theta}$ is sufficient for the precise model $(S_\theta)_{\theta\in\Theta}$. The following theorem states that these two models are even equivalent.

Theorem 3.27 *Let the index set Θ be finite with cardinality n. Let $\mathcal{X}$ be a set with σ-algebra $\mathcal{A}$ and $(P_\theta)_{\theta\in\Theta}$ a precise model where each P_θ is a measure; let $(S_\theta)_{\theta\in\Theta}$ be defined as above.*
Then, the precise models $(P_\theta)_{\theta\in\Theta}$ and $(S_\theta)_{\theta\in\Theta}$ are equivalent.

The content of Theorem 3.27 is well known; references for standard models are Blackwell (1951), Buja (1984), Strasser (1985), Le Cam (1986) and Torgersen (1991). In these references, standard models are called *standard experiments*. Though the content of Theorem 3.27 is well known and there are several references for standard models, it does not seem to be possible to satisfactorily cite a reference which is precisely in accordance with the definitions and the setup used in this book. Therefore, a self-contained proof of Theorem 3.27 is given below. This avoids involved conversions from the cited references and seems to be a convenient service for the readers.

Proof of Theorem 3.27: As stated above, the definition of $(S_\theta)_{\theta\in\Theta}$ immediately implies that $(P_\theta)_{\theta\in\Theta}$ is sufficient for $(S_\theta)_{\theta\in\Theta}$. Hence, it only remains to proof the converse direction.

To this end, recall the definitions of P, β, β_θ, $\mathcal{U}$, $\mathcal{C}$, S and ι_θ from the beginning of the present subsection.

Let $\mathbb{E}_P[f|\beta = u]$ be the conditional expectation of f given $\beta = u$ with respect to the probability measure P for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ and every $u \in \mathcal{U}$; cf. e.g. (Hoffmann-Jørgensen, 1994a, § 6.7) for conditional expectations. Let $\mathrm{ca}(\mathcal{U},\mathcal{C},S)$ be the set of all bounded charges which are absolutely continuous with respect to S; cf. (Bhaskara Rao and Bhaskara Rao, 1983, Definition 6.1.1). The "ca" in this notation is justified because every bounded charge $\mu \in \mathrm{ba}(\mathcal{U},\mathcal{C})$ which is absolutely continuous with respect to some $S \in \mathrm{ca}(\mathcal{U},\mathcal{C})$ is also an element of $\mathrm{ca}(\mathcal{U},\mathcal{C})$ according to (Bhaskara Rao and Bhaskara Rao, 1983, 6.1.11).

Put

$$\sigma(\mu)[f] \;:=\; \int \mathbb{E}_P[f|\beta = u]\,\mu(du) \qquad \forall\,\mu \in \mathrm{ca}(\mathcal{U},\mathcal{C},S)\,, \quad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}))$$

Though $u \mapsto \mathbb{E}_P[f|\beta = u]$ is only defined S-almost sure, $\sigma(\mu)[f]$ is defined well for every μ which is absolutely continuous with respect to S – i.e. for every $\mu \in \mathrm{ca}(\mathcal{U},\mathcal{C},S)$. Put $\sigma(\mu)\,:\,f \mapsto \sigma(\mu)[f]$. Then, it is easy to see that the properties of conditional expectations (Hoffmann-Jørgensen, 1994a, § 6.8) imply that

$$\sigma \;:\; \mathrm{ca}(\mathcal{U},\mathcal{C},S) \;\to\; \mathrm{ba}(\mathcal{X},\mathcal{A}))\,, \qquad \mu \;\mapsto\; \sigma(\mu)$$

is a transition in the sense of Definition 3.34. $\mathrm{ca}(\mathcal{U},\mathcal{C},S)$ is a band in $\mathrm{ba}(\mathcal{U},\mathcal{C})$ according to (Bhaskara Rao and Bhaskara Rao, 1983, Theorem 6.2.2) and Corollary 8.9. Hence, it follows from Lemma 8.30 that σ can be extended to a transition

$$\sigma \;:\; \mathrm{ba}(\mathcal{U},\mathcal{C}) \;\to\; \mathrm{ba}(\mathcal{X},\mathcal{A}))\,, \qquad \mu \;\mapsto\; \sigma(\mu) \tag{3.9}$$

[6] where $\beta(\mu)$ denotes the image measure defined by $\beta(\mu)(C) = \mu(\beta^{-1}(C))$

According to Proposition 3.36, this extended transition is a generalized randomization

$$\sigma \ : \ \mathrm{ba}(\mathcal{U},\mathcal{C}) \ \to \ \mathrm{ba}(\mathcal{X},\mathcal{A}))$$

Finally, for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}))$,

$$\begin{aligned}
\sigma(S_\theta)[f] \ &= \ \int \mathbb{E}_P[f|\beta = u]\, S_\theta(du) \ \overset{(3.8)}{=} \ \int \mathbb{E}_P[f|\beta = u] \cdot n\iota_\theta(u)\, S(du) \ = \\
&\overset{\text{(i)}}{=} \ n \cdot \int \mathbb{E}_P[f|\beta] \cdot (\iota_\theta \circ \beta)\, dP \ \overset{\text{(ii)}}{=} \ n \cdot \int \mathbb{E}_P\big[f \cdot (\iota_\theta \circ \beta) \,\big|\, \beta\,\big]\, dP \ = \\
&\overset{\text{(iii)}}{=} \ n \cdot \int f \cdot (\iota_\theta \circ \beta)\ dP \ = \ \int f\beta_\theta\, dP \ = \ \int f\, dP_\theta \ = \ P_\theta[f]
\end{aligned}$$

where (i) follows from the transformation theorem (Hoffmann-Jørgensen, 1994a, § 3.15), (ii) follows from (Hoffmann-Jørgensen, 1994a, (6.8.2)) and (iii) follows from (Hoffmann-Jørgensen, 1994a, (6.8.4)). That is, $\sigma(S_\theta) \ = \ P_\theta$ for every $\theta \in \Theta$ and, therefore, $(S_\theta)_{\theta\in\Theta}$ is sufficient for $(P_\theta)_{\theta\in\Theta}$. □

Remark 3.28 *The above presentation is similar to (Buja, 1984, § 5) and also the proof of Theorem 3.27 has some connections to (Buja, 1984, p. 374f) where an ordinary randomization plays the role of σ in (3.9). This ordinary randomization is given by a* regular *version of the conditional expectation, which is possible because $(\mathcal{X},\mathcal{A})$ is assumed to be a Polish space in Buja (1984). Such an assumption is not possible here because $(\mathcal{X},\mathcal{A})$ is given by a canonical Stone space below which – in general – is not a Polish space.*
Therefore, the conditional expectation cannot be assumed to define an ordinary randomization. Instead, we get a generalized randomization and have to use the theory of vector lattices in the proof of Theorem 3.27 in order to get around the problems caused by null sets in the definition of conditional expectations.

Now, we can define standard models for general precise models $(Q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$ where Q_θ does not have to be a probability measure (but a probability charge) and $\mathcal{B}$ does not have to be a σ-algebra (but an algebra).

Definition 3.29 *Let $(\mathcal{Y},\mathcal{B},Q_\theta)_{\theta\in\Theta})$ be a precise model. Let $(\mathcal{X},\mathcal{A}) \ = \ (\Xi,\mathfrak{B}_0)$ be the canonical Stone space and the precise model $(P_\theta)_{\theta\in\Theta}$ be given by*

$$P_\theta \ = \ \phi_0(Q_\theta)\,, \qquad \theta \in \Theta$$

where ϕ_0 is the canonical Stone transition. For $(P_\theta)_{\theta\in\Theta}$, let S and $(S_\theta)_{\theta\in\Theta}$ be defined as (3.6) and (3.7). Then, S is called standard measure of $(Q_\theta)_{\theta\in\Theta}$ *and $(\mathcal{U},\mathcal{C},(S_\theta)_{\theta\in\Theta}$ is called* standard model of $(\mathcal{Y},\mathcal{B},(Q_\theta)_{\theta\in\Theta})$.

Standard measure and standard model are defined well because $\mathfrak{B}_0$ is a σ-algebra and the canonical Stone transition defines probability measures $P_\theta \ = \ \phi_0(Q_\theta)$ according to Theorem 2.43.

Theorem 3.30 *Let $(\mathcal{Y},\mathcal{B},(Q_\theta)_{\theta\in\Theta})$ be a precise model and let $(\mathcal{U},\mathcal{C},(S_\theta)_{\theta\in\Theta})$ be its standard model.*
Then, $(Q_\theta)_{\theta\in\Theta}$ and $(S_\theta)_{\theta\in\Theta}$ are equivalent.

Proof: Let ϕ_0 be the canonical Stone transition and put

$$P_\theta \;=\; \phi_0(Q_\theta) \qquad \forall\,\theta\in\Theta$$

Q_θ and P_θ are probability charges and, therefore, they are also coherent upper expectations. According to Theorem 3.22, $(Q_\theta)_{\theta\in\Theta}$ and $(P_\theta)_{\theta\in\Theta}$ are equivalent (in the sense of Definition 3.17) as imprecise models. Hence, they are also equivalent (in the sense of Definition 3.14) as precise models; cf. Remark 3.21.
According to Theorem 3.27, $(P_\theta)_{\theta\in\Theta}$ and $(S_\theta)_{\theta\in\Theta}$ are equivalent.
Together, this implies that $(Q_\theta)_{\theta\in\Theta}$ and $(S_\theta)_{\theta\in\Theta}$ are equivalent. □

In the following, precise models $(Q_\theta)_{\theta\in\Theta}$ are sometimes abbreviated by calligraphical letters $\mathcal{F}$, i.e.

$$\mathcal{F} \;=\; (Q_\theta)_{\theta\in\Theta}$$

In this case, the standard measure of $\mathcal{F}=(Q_\theta)_{\theta\in\Theta}$ is denoted by

$$S^{\mathcal{F}}$$

and the standard model of $\mathcal{F}=(Q_\theta)_{\theta\in\Theta}$ is denoted by

$$(S_\theta^{\mathcal{F}})_{\theta\in\Theta}$$

Furthermore, if $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is a family of credal sets which corresponds to an imprecise model, the expression

$$\mathcal{F} \;\in\; (\mathcal{M}_\theta)_{\theta\in\Theta}$$

means that $\mathcal{F}$ denotes a precise model $\mathcal{F}=(Q_\theta)_{\theta\in\Theta}$ such that $Q_\theta\in\mathcal{M}_\theta$ for every $\theta\in\Theta$.

Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model and let $\mathcal{M}_\theta$ be the credal set of $\overline{Q}_\theta$ for every $\theta\in\Theta$.
Then, each precise model $\mathcal{F}=(Q_\theta)_{\theta\in\Theta}$ where $Q_\theta\in\mathcal{M}_\theta$ for every $\theta\in\Theta$ has a standard measure and a standard model. Now we can take the supremum over all standard measures so that we get a coherent upper prevision, which may be called "standard upper prevision". In the same way, we can take the supremum over all standard models so that we get an imprecise model, which may be called "standard imprecise model". This is the content of the following definition which is an analog to the corresponding definition in (Buja, 1984, § 5).

Definition 3.31 (Standard upper prevision, standard imprecise model)
Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model and let $\mathcal{M}_\theta$ be the credal set of $\overline{Q}_\theta$ for every $\theta\in\Theta$. For every $\mathcal{F}\in(\mathcal{M}_\theta)_{\theta\in\Theta}$, let $S^{\mathcal{F}}$ be the standard measure of $\mathcal{F}$ and $(S_\theta^{\mathcal{F}})_{\theta\in\Theta}$ be the standard model of $\mathcal{F}$.
Put

$$\overline{S}[h]=\sup\left\{S^{\mathcal{F}}[h]\;\middle|\;\mathcal{F}\in(\mathcal{M}_\theta)_{\theta\in\Theta}\right\}\qquad\forall\,h\in\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$$

$$\overline{S}_\theta[h]=\sup\left\{S_\theta^{\mathcal{F}}[h]\;\middle|\;\mathcal{F}\in(\mathcal{M}_\theta)_{\theta\in\Theta}\right\}\qquad\forall\,h\in\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$$

$\overline{S}$ *is called* standard upper prevision of $\overline{Q}_\theta)_{\theta\in\Theta}$ *and* $(\overline{S}_\theta)_{\theta\in\Theta}$ *is called* standard imprecise model of $(\overline{Q}_\theta)_{\theta\in\Theta}$.

Note that $\overline{S}$ is in fact a coherent upper prevision on $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ and $(\overline{S}_\theta)_{\theta\in\Theta}$ is in fact an imprecise model on $(\mathcal{U},\mathcal{C})$ – moreover, $\overline{S}$ and $\overline{S}_\theta$ are even upper expectations.

As stated in Section 2.5, the canonical Stone representation is interesting from a theoretical point of view because it enables us to work with *σ-additive* probability measures on σ-algebras whenever we like, i.e. a coherent upper prevision can always be represented by an upper expectation. However, we have to go over to the canonical stone space then which is a rather odd space. Theorem 3.30 shows that, at least in case of precise models, we are not tied up in the canonical Stone space; we may go over to standard models which are defined on $\mathcal{U}$ which is a very nice subset of $\mathbb{R}^n$. Especially, $\mathcal{U}$ is a compact Polish space. However, note that it has not been stated that an imprecise model and its imprecise standard model would be equivalent. This seems to be not true in general.

Summing up, standard models share two important properties: Firstly, they are defined on the very nice measurable space $(\mathcal{U},\mathcal{C})$. Secondly, they consist of linear previsions S_θ which are σ-additive probability measures. Furthermore, there is a standard model for every precise model and both models are equivalent. This is used in Subsection 4.2.1 where minimal Bayes risks are expressed in terms of (upper) standard measures.

Remark 3.32

a) *In the definition of standard measure and standard model, it is not possible to omit the intermediate step with the canonical Stone representation. This is due to the fact that the definition of S and S_θ in the beginning of the present subsection heavily rely on σ-additivity.*[7] *General precise models may consist of probability charges which lack σ-additivity. Therefore, we have to go over to the (somehow awkward) canonical Stone space in order to obtain probability measures. Next, we can define standard measures and standard models for these probability measures just as introduced in Blackwell (1951).*
Since Buja (1984) uses the classical setup consisting of probability measures on Polish spaces, the intermediate step with the canonical Stone representation is not needed there.

However, the usual definition of standard measures raises the following question: Let $(\mathcal{Y},\mathcal{B},(\overline{Q}_\theta)_{\theta\in\Theta})$ be an imprecise model such that $\mathcal{B}$ is a σ-algebra and each $\overline{Q}_\theta$ is a σ-additive probability measure on $\mathcal{B}$. Then, its standard measure could be defined in two different ways: Firstly, it could be defined via the canonical Stone space as done in Definition 3.31. Secondly, it could directly be defined via 3.6 without the intermediate step (with the canonical Stone space) as done in Blackwell (1951) and Buja (1984). Then, the question is: Do these two definitions coincide in this case? The answer is: Yes! This follows from part b) below.

b) *Let $(\mathcal{X}_1,\mathcal{A}_1,(P_{1,\theta})_{\theta\in\Theta})$ and $(\mathcal{X}_2,\mathcal{A}_2,(P_{2,\theta})_{\theta\in\Theta})$ be two precise models such that $\mathcal{A}_i$ is a σ-algebra and each $P_{i,\theta}$ is a probability measure on $\mathcal{A}_i$, $i\in\{1,2\}$.*

Then, $(P_{1,\theta})_{\theta\in\Theta}$ and $(P_{2,\theta})_{\theta\in\Theta}$ are equivalent, if and only if

$$S_1 \;=\; S_2$$

where S_i is the probability measure for $(P_{i,\theta})_{\theta\in\Theta}$ defined via (3.6).

[7] σ-additivity is necessary to ensure existence of β_θ according to the Radon-Nikodym-theorem.

The "only if" part follows from (3.8) and Theorem 3.27. For the proof of the "if" part, it is referred to Le Cam now:
Equivalence implies that the conical measures m_1 of $(P_{1,\theta})_{\theta\in\Theta}$ and m_2 of $(P_{2,\theta})_{\theta\in\Theta}$ coincides

$$m \;:=\; m_1 \;=\; m_2$$

according to (Le Cam, 1986, Theorem 2.4.4). S_1 and S_2 are localizations of m on $\mathcal{U}$ according to (Le Cam, 1986, p. 33). Hence, they coincide on Γ which denotes the set of all restrictions to $\mathcal{U}$ of the elements of the Choquet lattice $\mathcal{H}$. According to (Le Cam, 1986, p. 34), Γ is uniformly dense in $\mathcal{C}(\mathcal{U})$. This implies $S_1 = S_2$.

c) *Let $(\mathcal{X}_1, \mathcal{A}_1, (Q_{1,\theta})_{\theta\in\Theta})$ and $(\mathcal{X}_2, \mathcal{A}_2, (Q_{2,\theta})_{\theta\in\Theta})$ be precise models. Then, $(Q_{1,\theta})_{\theta\in\Theta}$ and $(Q_{2,\theta})_{\theta\in\Theta}$ are equivalent if and only if their standard measures coincide. (In contrast to part b, it is not assumed here that $Q_{1,\theta}$ and $Q_{2,\theta}$ would have to be σ-additive probability measures.)*

For the proof of this statement, let $(P_{1,\theta})_{\theta\in\Theta}$ and $(P_{2,\theta})_{\theta\in\Theta}$ denote their canonical Stone representations. Let S_1 and S_2 be their standard measures. According to Theorem 3.22, $(Q_{1,\theta})_{\theta\in\Theta}$ and $(Q_{2,\theta})_{\theta\in\Theta}$ are equivalent if and only if $(P_{1,\theta})_{\theta\in\Theta}$ and $(P_{2,\theta})_{\theta\in\Theta}$ are equivalent. $(P_{1,\theta})_{\theta\in\Theta}$ and $(P_{2,\theta})_{\theta\in\Theta}$ are equivalent if and only if $S_1 = S_2$ according to Definition 3.29 and part b of the present remark.

3.4 Connection to Le Cam's general setup

3.4.1 Outline of Le Cam's general setup

3.4.1.1 Introduction

As stated in Subsection 3.3.1, results from Le Cam (1964) and Le Cam (1986) can be used in this book because the definitions of precise models [8], (generalized) randomizations and sufficiency are in line with L. Le Cam's definitions. However, these results are formulated in a very abstract setup. The difficulty in reading Le Cam (1986) is also mentioned in the review (Strasser, 2008) of (Le Cam, 1986):

> "This book is not a text book for beginners. It is rather the master's report on his life's workshop of research. The author's style has been known for many years. From the reader he demands a lot. The reader has to be a connaisseur of classical statistics and argumentation. He must enjoy mathematics of any level of abstraction and sophistication. He must be willing to do his own proofs if the author considers them as not worth mentioning, which is not seldom the case. At the end the reader is rewarded by a host of ideas which is hard to match."

The bulk of the high level of abstraction comes from the fact that Le Cam usually does not use the measure theoretic formulation of probability theory and statistics. The traditional measure theoretic setup is based on a set of outcomes Ω and a σ-algebra $\mathcal{A}$ on Ω; probabilities are modeled by positive, normalized measures

$$P \;:\; \mathcal{A} \to \mathbb{R}$$

[8] called *experiments* in Le Cam (1986)

Le Cam's abstract setup dispenses with the sample space $(\Omega, \mathcal{A})$ and, consequently, probabilities cannot be defined by measures. Instead probabilities are defined to be the positive, normalized elements of abstract L-spaces. This proceeding has some advantages and does essentially not differ from the traditional definitions but is on a rather high level of abstraction. Unfortunately, the connections between this abstract setup and the traditional measure theoretic setup is not explained adequately in Le Cam (1986) – this is the main reason why Le Cam (1986) is hard to read and why it is not appropriate to become familiar with L. Le Cam's seminal ideas and concepts.

So, the present subsection has two aims: Firstly, it is intended to be a comprehensible introduction in L. Le Cam's abstract setup for those who wants to become familiar with it. Secondly, it is shown that the decision theoretic definitions of the present book are in line with the definitions in Le Cam (1964) and Le Cam (1986).
Before continuing with the present subsection, it may me helpful to read (van der Vaart, 2002, § 7 and § 8) which is an excellent outline of the abstract setup. Furthermore, some basics of the theory of vector lattices are needed; these are recollected in Subsection 8.1. An interesting article about the life of L. Le Cam is (Yang, 2002).

After the above paragraphs which emphasized the difficulties which are connected with the study of Le Cam (1986), the following citation from the review (Strasser, 2008) of Le Cam (1986) may indicate why this is worthwhile:

> "Until now, only a very little part of the author's work has found its way into the international research business, and this little part brought forth cascades of successor papers (on 'contiguity', on 'three lemmas of LeCam', on 'the asymptotic minimax bound', etc.). The remaining 95 percent of the book will keep people busy for decades."

In the beginning of the 20th century, it was a demanding task to find a mathematically rigorous way to model probabilities. This task took several decades and some of the best mathematicians worked on it. Not until 1933, the discovery was published by A.N. Kologorov[9] that measure theory is quite suitable to model probabilities. This was the starting point of the mathematical theory of probability and statistics. Still in 1929, Bertrand Russell remarked in a lecture:

> "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means."[10]

In spite of these difficulties, many modern textbooks on probability theory gives the impression that the measure theoretic formulation of probabilities would have been obvious.[11] Furthermore, it is hardly explained what the sample space $(\Omega, \mathcal{A})$ really means. In order to understand why it is possible to dispense with the sample space in the general setup, it is helpful to recall the meaning of the sample space at first.

3.4.1.2 The sample space $(\Omega, \mathcal{A})$

Usually, a statistical evaluation starts with the fixing of a sample space $(\Omega, \mathcal{A})$. Here, Ω is a set where each element $\omega \in \Omega$ represents a possible outcome of a random experiment.

[9] Kolmogoroff (1933)

[10] Confer (Bell, 1992, p. 587).

[11] A nice exception is e.g. Hoffmann-Jørgensen (1994a).

So each subset $A \subset \Omega$ represents a whole set of possible outcomes – such sets A are called *events*. $\mathcal{A}$ is a subset of the power set 2^Ω of Ω

$$\mathcal{A} \subset 2^\Omega$$

Here, $\mathcal{A}$ represents the collection of all those events $A \subset \Omega$ which can be observed in principle. That is, the events $A \in \mathcal{A}$ are exactly those events where the experimenter can decide if it has occurred or not. As an example, assume that the outcome of an experiment may be any real number $\omega \in \Omega = \mathbb{R}$, e.g.

$$\omega = 26.5269725476391073785306636185478\ldots$$

However, the accuracy of the measurement is usually limited. So, assume for example that the measuring instrument is accurate up to the second decimal place. That is, the experimenter can only decide in which interval

$$\big(\tfrac{k}{100} - 0.005 \; , \; \tfrac{k}{100} + 0.005\big] \; , \qquad k \in \mathbb{Z}$$

the outcome ω lies. Therefore, the observable events have the following form

$$A = \bigcup_{k \in K} \big(\tfrac{k}{100} - 0.005 \; , \; \tfrac{k}{100} + 0.005\big] \; , \qquad K \subset \mathbb{Z}$$

i.e.

$$\mathcal{A} = \left\{ \bigcup_{k \in K} \big(\tfrac{k}{100} - 0.005 \; , \; \tfrac{k}{100} + 0.005\big] \;\middle|\; K \subset \mathbb{Z} \right\}$$

Why is the collection of the observable events $\mathcal{A}$ usually assumed to be a σ-algebra? This gets clear by the above interpretation of the events $A \in \mathcal{A}$: *The events $A \in \mathcal{A}$ are all those events where we can decide if they have occurred or not.*
Therefore,

$$\emptyset \in \mathcal{A} \qquad \text{and} \qquad \Omega \in \mathcal{A} \tag{3.10}$$

because we can decide if the impossible event $\emptyset$ has occurred (it has not occurred, of course) and we can decide if the certain event Ω has occurred (it must have occurred, of course).
Next, we have

$$A \in \mathcal{A} \qquad \Rightarrow \qquad A^{\mathrm{C}} \in \mathcal{A} \tag{3.11}$$

because, if we can decide if A has occurred, then we can also decide if the complement $A^{\mathrm{C}} = \Omega \setminus A$ has occurred:

$$\begin{aligned} A \text{ has occurred} \quad &\Rightarrow \quad A^{\mathrm{C}} \text{ has not occurred} \\ A \text{ has not occurred} \quad &\Rightarrow \quad A^{\mathrm{C}} \text{ has occurred} \end{aligned}$$

Furthermore,

$$A_1, A_2, \ldots, A_n \in \mathcal{A} \qquad \Rightarrow \qquad \bigcup_{k=1}^{n} A_k \in \mathcal{A} \tag{3.12}$$

because, if we can decide for every $k \in \{1,\dots,n\}$ if A_k has occurred, then we can decide if $\bigcup_{k=1}^n A_k$ has occurred:

$$\begin{array}{ll} \displaystyle\bigcup_{k=1}^{n} A_k \text{ has occurred} & \text{if there is at least one } A_k \text{ which has occurred.} \\ \displaystyle\bigcup_{k=1}^{n} A_k \text{ has not occurred} & \text{if none of the } A_k \text{ has occurred.} \end{array}$$

Summing up, (3.10), (3.11) and (3.12) implies that $\mathcal{A}$ has to be an algebra, at least. Usually, it is assumed that $\mathcal{A}$ is a σ-algebra – that is, (3.12) is slightly strengthened.

The above description of $\mathcal{A}$ can also be found in (Hoffmann-Jørgensen, 1994a, § 6.1). There, $\mathcal{A}$ is also called *information* and the following figurative explanation is given:

> "Information can also be described as a net on Ω such that two outcomes in a mesh cannot be distinguished by the information available, but outcomes in two different meshes can. You may think of such nets as a map. On a world map, it is not possible to distinguish the Empire State Building and the United Nation Building in New York City, but a map of New York contains enough information to distinguish the two sites." (Hoffmann-Jørgensen, 1994a, p. 441)

On the one hand, this setup based on sample spaces is rather descriptive but, on the other hand, it raises some difficulties as has been pointed out by (Le Cam, 1986, § 1.1): The set Ω represents the results of the experiments – however, the choice of the set Ω is usually rather arbitrary. For instance, we have chosen $\Omega = \mathbb{R}$ in the above example but it would also have been possible to chose a suitable sample space where $\Omega = \mathbb{Z}$. Of course, $\mathbb{R}$ and $\mathbb{Z}$ are quite different sets; $\mathbb{Z}$ is a discrete, countable set whereas none of this is true for $\mathbb{R}$. In order to avoid consequences which depend on the choice of Ω, it would be better to dispense with Ω. However, $\mathcal{A}$ depends on Ω: If we choose $\Omega = \mathbb{Z}$, we would get $\mathcal{A} = 2^{\mathbb{Z}}$. So, if we dispense with Ω we will also have to dispense with $\mathcal{A}$.
Of course, both choices of $(\Omega, \mathcal{A})$ in the above example *essentially* lead to the same random variables: That is the sets

$$\mathcal{L}_\infty(\Omega, \mathcal{A})$$

are isomorphic as M-spaces for the different choices of $(\Omega, \mathcal{A})$. This M-space structure (which is preserved by isomorphisms) contains the essential structure of the sample space. (Le Cam, 1986, p. 3) argues:

> "Let us (...) agree that there are certain objects called 'random variables' which have a life of their own in the physical world but have the property that if 'measured' in an experiment they produce a real number. (...) If two 'random variables' can be 'measured' in the same experiment, one obtains two numbers which can be added, multiplied, etc. One can also take the *minimum* or *maximum* of the two, multiply them by other real numbers, and so forth."

In this way, the experiment should not be represented by $(\Omega, \mathcal{A})$ but by the M-space structure of $\mathcal{L}_\infty(\Omega, \mathcal{A})$. The following paragraph describes how this can be done.

3.4.1.3 Dispensing with the sample space

According to the above reasoning, the M-space structure of $\mathcal{L}_\infty(\Omega,\mathcal{A})$ contains the essential information about the sample space. So, we may "forget" any additional structure of $\mathcal{L}_\infty(\Omega,\mathcal{A})$ which goes beyond its M-space structure. If we do this and consider $\mathrm{ba}(\Omega,\mathcal{A})$ as the dual of $\mathcal{L}_\infty(\Omega,\mathcal{A})$, nothing remains left from $\mathrm{ba}(\Omega,\mathcal{A})$ apart from its L-space structure (remember that the dual space of an M-space is always an L-space; cf. Proposition 8.22 a)). That is, the L-space structure contains the essential information about the elements of $\mathrm{ba}(\Omega,\mathcal{A})$. In terms of L-spaces, the probability charges $P \in \mathrm{ba}(\Omega,\mathcal{A})$ are precisely the normed, positive elements of $\mathrm{ba}(\Omega,\mathcal{A})$ (cf. Subsection 2.2) – that is, the *probability* charges can completely be identified by the L-space structure of $\mathrm{ba}(\Omega,\mathcal{A})$!

If we have a fixed precise model

$$\mathcal{E} \;=\; (P_\theta)_{\theta\in\Theta} \;\subset\; \mathrm{ba}_1^+(\Omega,\mathcal{A})$$

we usually do not have to consider the whole L-space $\mathrm{ba}(\Omega,\mathcal{A})$ but it is enough to consider the smallest L-space $L \subset \mathrm{ba}(\Omega,\mathcal{A})$ which contains our model $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$. This L-space L is equal to the smallest band in the L-space $\mathrm{ba}(\Omega,\mathcal{A})$ which contains $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$. This set L is called *L-space of* $\mathcal{E}$ or *L-space generated by* $\mathcal{E}$ and is denoted by

$$L(\mathcal{E})$$

in Le Cam (1986).

In classical statistics, $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ is a family of probability measures which is dominated[12] by some σ-finite measure μ. In this case,

$$L(\mathcal{E}) \;\subset\; \big\{\nu \in \mathrm{ba}(\Omega,\mathcal{A}) \;\big|\; d\nu = f\,d\mu\,,\;\; f \in L_1(\Omega,\mathcal{A},\mu)\big\}$$

In addition, assume that μ is also dominated[13] by $\mathcal{E}$. Then,

$$L(\mathcal{E}) \;=\; \big\{\nu \in \mathrm{ba}(\Omega,\mathcal{A}) \;\big|\; d\nu = f\,d\mu\,,\;\; f \in L_1(\Omega,\mathcal{A},\mu)\big\} \tag{3.13}$$

That is, $L(\mathcal{E})$ can be identified with the set of densities $L_1(\Omega,\mathcal{A},\mu)$ then. Examples for this case are

- models $(P_\theta)_{\theta\in\Theta}$ on $(\mathbb{Z},2^{\mathbb{Z}})$
- the model $(P_\theta)_{\theta\in\Theta}$ on $(\mathbb{R},\mathbb{B})$ where $P_\theta = \mathcal{N}(a,b)$ for every $(a,b) = \theta \in \Theta$

Note that

$$L(\mathcal{E}) \;=\; \mathrm{ba}(\Omega,\mathcal{A}) \tag{3.14}$$

is also possible for a suitable chosen model $\mathcal{E}$.[14]

So far, we have stated that the L-space structure of $L(\mathcal{E})$ contains all essential information about $\mathcal{E}$ – that is, it is not important that the elements P_θ of $\mathcal{E}$ are probability measures on some specific sample space $(\Omega,\mathcal{A})$ and they may be defined as elements of any abstract L-space.

Accordingly, L. Le Cam proposes the following definition of experiments – called precise models in the present book. In order to avoid confusions with the definitions given in previous sections, the definitions given by L. Le Cam in his general setup carry the prefix "LC".

[12] That is: $\mu(A) = 0\,,\;\; A \in \mathcal{A} \quad\Rightarrow\quad P_\theta(A) = 0 \;\;\forall\,\theta \in \Theta$

[13] That is: $P_\theta(A) = 0 \;\;\forall\,\theta \in \Theta\,,\;\; A \in \mathcal{A} \quad\Rightarrow\quad \mu(A) = 0$

[14] Choose $\Theta = \mathrm{ba}_1^+(\Omega,\mathcal{A})$ and $P_\theta = \theta \;\;\forall\,\theta \in \Theta\,.$

Definition 3.33 (Experiment / precise model)
An LC-experiment /precise LC-model indexed by the set Θ *is a family* $(P_\theta)_{\theta\in\Theta} \subset L$ *where* L *is an L-space and* P_θ *is a normed* ($\|P_\theta\| = 1$)*, positive* ($P_\theta \geq 0$) *element of* L *for every* $\theta \in \Theta$ *.*
Cf. (Le Cam, 1986, p. 5).

It is clear that every ordinary precise model (according to Section 3.2) is a precise LC-model
Of course, the question arises if the above definition excessively generalizes the notion "precise model". As described below, the answer is: no, essentially not! Indeed, it is less a generalization then an abstraction.

Since we have lost the sample space $(\Omega, \mathcal{A})$ now, we have also lost the random variables $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$. In order to reintroduce them, we consider the dual space of L denoted by

$$L^* \;=:\; M$$

The elements $f^* \in M$ corresponds to the random variables $f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$.

For example, let the precise model $\mathcal{E} = (P_\theta)_\theta$ consist of probability measures P_θ on a sample space $(\Omega, \mathcal{A})$ such that P_θ is dominated by a σ-finite measure μ on $(\Omega, \mathcal{A})$ for every $\theta \in \Theta$. Furthermore, assume that μ is also dominated by $\mathcal{E}$. Put $L = L(\mathcal{E})$; according to (3.13), we can identify $L(\mathcal{E})$ with the set of all μ-densities. Next, (Dunford and Schwartz, 1958, Theorem IV.8.5) says that the dual space

$$L^* = M$$

is equal to $L_\infty(\Omega, \mathcal{A}, \mu)$. Here, corresponding elements $f^* \in M$ and $f \in L_\infty(\Omega, \mathcal{A}, \mu)$ are related by the identity

$$f^*(\beta) \;=\; \int_\Omega f(\omega)\beta(\omega)\,\mu(d\omega) \qquad \forall\, \beta \in L_1(\Omega, \mathcal{A}, \mu)$$

That is, in this special case which is quite common in classical statistics, $L(\mathcal{E})$ is equal to the set of all μ-densities and M is equal to the set of all bounded random variables. The dual space of $L = L(\mathcal{E})$ is called *M-space of* $\mathcal{E}$ or *M-space generated by* $\mathcal{E}$ and is denoted by

$$M \;=\; M(\mathcal{E})$$

in Le Cam (1986). This is indeed an M-space because $L(\mathcal{E})$ is an L-space; cf. Section 8.1. The following schema illustrates the relations between the abstract setup and the traditional concepts:

bounded random variables $\quad L_\infty(\Omega, \mathcal{A}, \mu) \xrightarrow{\text{abstraction}} M(\mathcal{E})$

μ-densities $\quad L_1(\Omega, \mathcal{A}, \mu) \xrightarrow{\text{abstraction}} L(\mathcal{E})$

In order to describe a decision problem now, we need a set $\mathbb{D}$ of possible decisions $t \in \mathbb{D}$ and a loss function

$$\Theta \times \mathbb{D} \;\to\; \mathbb{R}\,, \qquad (\theta, t) \;\mapsto\; W_\theta(t)$$

In Le Cam's setup, we have

$$W_\theta \ \in \ \Gamma \qquad \forall\, \theta \in \Theta$$

where Γ is a set of bounded functions

$$\gamma : \ \mathbb{D} \ \to \ \mathbb{R}$$

which fulfills certain conditions (Γ is a so-called *uniform lattice*; cf. (Le Cam, 1986, 4 and 5)). As a special case, we may simply take

$$\Gamma \ = \ \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$$

where $\mathcal{D}$ is an algebra on $\mathbb{D}$.

Finally, Markov kernels play an important role in decision theory – especially as randomized decision procedures. Since Le Cam dispenses with sample spaces in the definition of LC-experiments / precise LC-models, Markov kernels cannot be defined. Therefore, Markov kernels are replaced by transitions:

Definition 3.34 *Let L_1 and L_2 be L-spaces. A* transition from L_1 to L_2 *is a map*

$$\sigma : \ L_1 \ \to \ L_2$$

which is

- *linear*
- *positive:* $T(\mu) \geq 0 \quad \forall\, \mu \geq 0$
- *normalized:* $\|T(\mu)\| = \|\mu\| \quad \forall\, \mu \geq 0$

Now, let Θ be an index set, L an L-space and

$$\mathcal{E} \ = \ (P_\theta)_{\theta\in\Theta} \ \subset \ L$$

a precise LC-model; $L(\mathcal{E})$ denotes the L-space generated by $\mathcal{E}$.

$$\Theta \times \mathbb{D} \ \to \ \mathbb{R}\,, \qquad (\theta, t) \ \mapsto \ W_\theta(t)$$

is a loss function where

$$W_\theta \ \in \ \Gamma := \mathcal{L}_\infty(\mathbb{D}, \mathcal{D}) \qquad \forall\, \theta \in \Theta$$

According to Theorem 2.4, the dual space of $\Gamma = \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ is equal to

$$\Gamma^* \ = \ \mathrm{ba}(\mathbb{D}, \mathcal{D})$$

which is an L-space; cf. Theorem 2.6. With these predefinitions, decision procedures in the sense of (Le Cam, 1986, § 1.3) can be defined:

Definition 3.35 (LC - decision procedures) *A* LC - decision procedure (based on $\mathcal{E}$ and taking values in $\mathbb{D}$) *is a transition from $L(\mathcal{E})$ to Γ^**

$$\sigma : \ L(\mathcal{E}) \ \to \ \Gamma^*$$

Next, the *LC - risk function* is defined to be

$$\Theta \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; \sigma(P_\theta)[W_\theta]$$

Since $\Gamma^* = \mathrm{ba}(\mathbb{D}, \mathcal{D})$, we may also write

$$\sigma(P_\theta)[W_\theta] \;=\; \int W_\theta(t)\, K_\theta(dt)\,, \qquad \text{where} \quad K_\theta := \sigma(P_\theta) \tag{3.15}$$

Instead of (3.15), Le Cam (1986) uses the notation

$$W_\theta \sigma P_\theta \;:=\; \sigma(P_\theta)[W_\theta]$$

How do these definitions fit into the usual setup based on sample spaces? In order to answer this question, let the precise model $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ consist of probability measures P_θ on a sample space $(\Omega, \mathcal{A})$ such that P_θ is dominated by a σ-finite measure μ on $(\Omega, \mathcal{A})$ for every $\theta \in \Theta$. Furthermore, assume that μ is also dominated by $\mathcal{E}$. Put $L = L(\mathcal{E})$ and $M = M(\mathcal{E})$. As stated above, we can identify

$$L(\mathcal{E}) \;=\; L_1(\Omega, \mathcal{A}, \mu)$$

and

$$M(\mathcal{E}) \;=\; L_\infty(\Omega, \mathcal{A}, \mu)$$

As usual, $(\mathbb{D}, \mathcal{D})$ is a decision space and the loss function is some

$$(W_\theta)_{\theta\in\Theta} \;\subset\; \mathcal{L}_\infty(\mathbb{D}, \mathcal{D}) \;=:\; \Gamma$$

Let τ be an ordinary (randomized) decision function, i.e. τ is a Markov kernel

$$\tau :\;\; \Omega \times \mathcal{D} \;\to\; \mathbb{R}\,, \qquad (\omega, D) \;\mapsto\; \tau_\omega(D)$$

The Markov kernel τ defines a transition

$$\sigma :\;\; L(\mathcal{E}) \;\to\; \mathrm{ba}(\mathbb{D}, \mathcal{D}) \;=\; \Gamma^*$$

via

$$\sigma(\nu)[h] \;=\; \int_\Omega \int_{\mathbb{D}} h(t)\, \tau_\omega(dt)\, \nu(d\omega) \;=\; \int_\Omega \int_{\mathbb{D}} h(t) \beta(\omega)\, \tau_\omega(dt)\, \mu(d\omega)$$

for every $h \in \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ and every $d\nu = \beta\, d\mu$, $\beta \in L_1(\Omega, \mathcal{A}, \mu)$.
Therefore, the risk function is equal to

$$\theta \;\mapsto\; W_\theta \sigma P_\theta \;=\; \sigma(P_\theta)[W_\theta] \;=\; \int_\Omega \int_{\mathbb{D}} W_\theta(t)\, \tau_\omega(dt)\, P_\theta(d\omega)$$

In this way, LC - decision procedures generalize ordinary (randomized) decision functions. Again, the question arises if this concept is an *excessive* generalization. In the following, it is explained why the answer to this question is "no, essentially not".

3.4.1.4 Concrete representations of the general concepts

Now, it is explained why Le Cam's definitions are rather abstractions than generalizations of the ordinary definitions: As a matter of fact, there is always a suitable measurable space $(\Xi, \mathfrak{B}_0)$ and a suitable decision space $(\hat{\mathbb{D}}, \hat{\mathcal{D}})$ such that the L-space

$$L(\mathcal{E}) \qquad \text{may be represented by a subset of} \qquad \mathrm{ca}(\Xi, \mathfrak{B}_0)$$

the M-space

$$M(\mathcal{E}) \qquad \text{may be represented by} \qquad \mathcal{C}(\Xi)$$

and every LC - decision procedure

$$\sigma \qquad \text{may be represented by a Markov kernel} \qquad \tau$$

Of course, these representations are interesting rather from a theoretical point of view than from a practical point of view. The reader who is not interested in such representations may skip this paragraph, which is based on (Le Cam, 1986, p. 12).

In order to find representations, we have to start with $M(\mathcal{E})$. This set $M(\mathcal{E})$ is the dual space of the L-space $L(\mathcal{E})$. Therefore, it is an M-space with unit (cf. (Schaefer, 1974, Proposition 9.1)). Next, there is a compact set Ξ such that

$$M(\mathcal{E}) \qquad \text{and} \qquad \mathcal{C}(\Xi)$$

are M-space isomorphic – this is the content of a famous theorem due to S. Kakutani, M. Krein and S. Krein (cf. (Schaefer, 1974, Theorem 7.4). Note that this theorem not only states existence but also specifies a concrete isomorphism (we may omit the explicit description of this isomorphism here). As a consequence, $M(\mathcal{E})$ may be represented as $\mathcal{C}(\Xi)$ because the M-space structures of these two sets coincide – remember that the M-space structure contains the essential information about the random variables. In this way, the elements of $L(\mathcal{E})$ corresponds to bounded linear functionals on $\mathcal{C}(\Xi)$ and the Daniell-Stone extension theorem (Dudley, 1989, § 4.5) implies that these bounded linear functionals may be represented by bounded signed measures $\mu \in \mathrm{ca}(\Xi, \mathfrak{B}_0)$; $\mathfrak{B}_0$ denotes the Baire-σ-algebra.
Since $\Gamma = \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ is also an M-space with unit, there is a measurable space $(\hat{\mathbb{D}}, \hat{\mathcal{D}})$ such that $\hat{\mathbb{D}}$ is compact, $\hat{\mathcal{D}}$ is the Baire-σ-algebra and

$$\Gamma = \mathcal{L}_\infty(\mathbb{D}, \mathcal{D}) \qquad \text{and} \qquad \mathcal{C}(\hat{\mathbb{D}})$$

are M-space isomorphic. That is, every loss function $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ may be represented as a loss function

$$(\hat{W}_\theta)_{\theta\in\Theta} \subset \mathcal{C}(\hat{\mathbb{D}}) \ =: \ \hat{\Gamma}$$

Now, let $\sigma : \ L(\mathcal{E}) \to \Gamma^*$ be an LC - decision procedure. Its adjoint is a map

$$\Gamma \ \to \ M(\mathcal{E}) \, , \qquad h \ \mapsto \ \sigma(\cdot)[h]$$

which may be represented as a map

$$\mathcal{C}(\hat{\mathbb{D}}) \ \to \ \mathcal{C}(\Xi) \, , \qquad \hat{h} \ \mapsto \ \varphi\Big(\sigma(\cdot)\big[\psi(\hat{h})\big]\Big)$$

where $\varphi : M(\mathcal{E}) \to \mathcal{C}(\Xi)$ and $\psi : \mathcal{C}(\hat{\mathbb{D}}) \to \mathcal{C}(\mathbb{D})$ are M-space isomorphisms. According to the Daniell-Stone theorem (Dudley, 1989, § 4.5) again, there is a unique bounded signed measure

$$\tau_x : \ \mathcal{L}_\infty(\hat{\mathbb{D}}, \hat{\mathcal{D}}) \ \to \ \mathbb{R}\,, \qquad \hat{h} \ \mapsto \ \tau_x[\hat{h}] \ = \ \varphi\Big(\sigma(\cdot)\big[\psi(\hat{h})\big]\Big)(x)$$

for every $x \in \Xi$. That is, the LC - decision procedure σ can be represented by the Markov kernel

$$\tau : \ \Xi \times \hat{\mathcal{D}} \ \to \ \mathbb{R}\,, \qquad (x, \hat{D}) \ \mapsto \ \tau_x(\hat{D})$$

by use of the M-space isomorphisms φ and ψ .

3.4.1.5 Advantages of the abstract setup

Obviously, the abstract setup has the disadvantage that the definitions are hardly intuitionally understandable and less appropriate for practical purposes.
However, these definitions are very suitable for general theoretical investigations. As described before, sample spaces represent the observable events and random variables $X : \Omega \to \mathbb{R}$ represent observations. As a matter of fact, different representations are usually possible for the same situation in the real world. Unfortunately, fixing a concrete sample space may always artificially generate problems of a measure theoretic type which are rather meaningless in the real world.

Accordingly, L. Le Cam argues

> "The point is that the above representation is very special and the usual setup where $\mathcal{E} = \{P_\theta,\, \theta \in \Theta\}$ is given by probability measures P_θ on a σ-field carried by a set $\mathcal{X}$ does not insure that the σ-field $\mathcal{A}$ or the set $\mathcal{X}$ are selected well enough to be able to proceed without trouble. The abstract framework avoids the troubles caused by $\mathcal{X}$ or similar sets by ignoring them." (Le Cam, 1986, p. 12)

and (van der Vaart, 2002, p. 662) further explains

> "While Le Cam would acknowledge a role for measure theory, his main objection to the usual way of describing statistical experiments is that a given practical situation might be describable by many different types of sample spaces and 'true' measures. If one happened to choose the 'wrong one', one might get burdened by technical problems, for no good reason. Furthermore, any experiment in Le Cam's sense can be represented as an experiment in the usual way if the sample space is chosen appropriately (. . .)."

3.4.2 Accordance with Le Cam's general setup

In the present section, it is shown that the decision theoretic definitions of the present book are in line with the definitions in Le Cam (1964) and Le Cam (1986).

To this end, recall from Theorem 2.6 that ba$(\Omega, \mathcal{A})$ is an L-space. The following Proposition states that the generalized randomizations (used in the present book) coincide with the transitions between these L-spaces.

Proposition 3.36 *Let Ω_1 and Ω_2 be sets with algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. Then, the generalized randomizations*

$$\sigma : \ \mathrm{ba}(\Omega_1,\mathcal{A}_1) \ \rightarrow \ \mathrm{ba}(\Omega_2,\mathcal{A}_2)$$

are precisely the transitions from $\mathrm{ba}(\Omega_1,\mathcal{A}_1)$ *to* $\mathrm{ba}(\Omega_2,\mathcal{A}_2)$.

Proof: It only remains to show that the two conditions of normalization coincide – i.e.

$$\|\mu\| \ = \ \mu\big[I_\Omega\big] \qquad \forall\, \mu \,\geq\, 0\,, \quad \mu \in \mathrm{ba}(\Omega,\mathcal{A})$$

However, this is the content of (2.10). □

Accordingly, the following proposition states that the restricted randomizations coincide with the finitely supported (Γ, H) continuous transitions. This is important in the proof of Theorem 3.10.
Here, we have

$$\Gamma := \mathcal{L}_\infty(\Omega_2,\mathcal{A}_2) \qquad \text{and} \qquad H := \mathcal{L}_\infty(\Omega_1,\mathcal{A}_1) \ \subset \ \big(\mathrm{ba}(\Omega_1,\mathcal{A}_1)\big)^*$$

Then, (Γ, H) continuity of a transition

$$\sigma : \ \mathrm{ba}(\Omega_1,\mathcal{A}_1) \ \rightarrow \ \mathrm{ba}(\Omega_2,\mathcal{A}_2)$$

is defined to be

> *For every $f_2 \in \mathcal{L}_\infty(\Omega_2,\mathcal{A}_2)$ there is a $T(f_2) = f_1 \in \mathcal{L}_\infty(\Omega_1,\mathcal{A}_1)$ such that*
>
> $$\sigma(\mu_1)[f_2] \ = \ \mu_1[f_1] \qquad \forall\, \mu_1 \in \mathrm{ba}(\Omega_1,\mathcal{A}_1) \tag{3.16}$$

by (Le Cam, 1986, p. 6). Furthermore, a transition $\sigma : \mathrm{ba}(\Omega_1,\mathcal{A}_1) \rightarrow \mathrm{ba}(\Omega_2,\mathcal{A}_2)$ is called finitely supported if

> *there is a fixed finite subset $\tilde{\Omega}_2 \subset \Omega_2$ such that for every $\mu_1 \in \mathrm{ba}(\Omega_1,\mathcal{A}_1)$ there are real numbers $\alpha_{\tilde{\omega}_2} \in \mathbb{R}$, $\tilde{\omega}_2 \in \tilde{\Omega}_2$, such that*
>
> $$\sigma(\mu_1)[f_2] \ = \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2} f_2(\tilde{\omega}_2) \qquad \forall\, f_2 \in \mathcal{L}_\infty(\Omega_2,\mathcal{A}_2) \tag{3.17}$$

cf. (Le Cam, 1986, p. 6).

Proposition 3.37 *Let Ω_1 and Ω_2 be sets with algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. Then, the restricted randomizations*

$$\sigma : \ \mathrm{ba}(\Omega_1,\mathcal{A}_1) \ \rightarrow \ \mathrm{ba}(\Omega_2,\mathcal{A}_2)$$

are precisely the finitely supported (Γ, H) *continuous transitions from* $\mathrm{ba}(\Omega_1,\mathcal{A}_1)$ *to* $\mathrm{ba}(\Omega_2,\mathcal{A}_2)$ *where*

$$\Gamma := \mathcal{L}_\infty(\Omega_2,\mathcal{A}_2) \qquad \textit{and} \qquad H := \mathcal{L}_\infty(\Omega_1,\mathcal{A}_1) \ \subset \ \big(\mathrm{ba}(\Omega_1,\mathcal{A}_1)\big)^*$$

Proof: It is a direct consequence of the definitions and Proposition 3.36 that every restricted randomization is a finitely supported (Γ, H) continuous transition.

Conversely, let

$$\sigma : \ \mathrm{ba}(\Omega_1, \mathcal{A}_1) \ \to \ \mathrm{ba}(\Omega_2, \mathcal{A}_2)$$

be a finitely supported (Γ, H) continuous transition. Then, Proposition 3.36 states that σ is a generalized randomization and (3.16) implies that for every $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$ there is a $T(f_2) \in \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$ such that

$$\sigma(\mu_1)[f_2] \ = \ \mu_1\big[T(f_2)\big] \qquad \forall\, \mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1) \tag{3.18}$$

Since $T(f_2)$ is uniquely determined by this equation for every $f_2 \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2)$, it is easy to see that the map

$$T : \ \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \ \to \ \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)\,, \qquad f_2 \ \mapsto \ T(f_2)$$

has the same properties as in Proposition 3.11 b). Then, it follows from Proposition 3.11 that σ is an ordinary randomization. That is, there is a finitely additive Markov kernel

$$\tau : \ \Omega_1 \times \mathcal{A}_2 \ \to \ \mathbb{R}\,, \qquad (\omega_1, A_2) \ \mapsto \ \tau_{\omega_1}(A_2)$$

such that

$$\sigma(\mu_1)[I_{A_2}] \ = \ \int \tau_{\omega_1}(I_{A_2})\, \mu_1(d\omega_1) \tag{3.19}$$

for every $\mu_1 \in \mathrm{ba}(\Omega_1, \mathcal{A}_1)$ and $A_2 \in \mathcal{A}_2$.

Next, (3.17) states that, there is a fixed finite subset $\tilde{\Omega}_2 \subset \Omega_2$ such that the following is true:

For every $\omega_1 \in \Omega_1$, there are real numbers $\alpha_{\tilde{\omega}_2}(\omega_1) \in \mathbb{R}$, $\tilde{\omega}_2 \in \tilde{\Omega}_2$, such that

$$\begin{aligned} \tau_{\omega_1}(A_2) \quad &= \quad \int \tau_{\omega_1}(A_2)\, \delta_{\omega_1}(d\hat{\omega}_1) \ \overset{(3.19)}{=} \ \sigma(\delta_{\omega_1})[I_{A_2}] \ = \\ &\overset{(3.17)}{=} \ \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2}(\omega_1) I_{A_2}(\tilde{\omega}_2) \ = \ \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2}(\omega_1) \delta_{\tilde{\omega}_2}(A_2) \end{aligned} \tag{3.20}$$

for every $A_2 \in \mathcal{A}_2$. Put $\alpha_{\tilde{\omega}_2} : \Omega_1 \ \to \ \mathbb{R}\,,\ \omega_1 \ \mapsto \ \alpha_{\tilde{\omega}_2}(\omega_1)$ for every $\tilde{\omega}_2 \in \tilde{\Omega}_2$.

In order to finish the proof, it remains to show that the maps $\alpha_{\tilde{\omega}_2}$ fulfill certain conditions which are listed in Definition 3.7. To this end, note that

$$T(I_{A_2})(\omega_1) \ \overset{(3.18)}{=} \ \sigma(\delta_{\omega_1})[I_{A_2}] \ \overset{(3.19\,,\,3.20)}{=} \ \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2}(\omega_1) \delta_{\tilde{\omega}_2}(A_2) \tag{3.21}$$

for every $\omega_1 \in \Omega_1$.

Firstly, fix some $\tilde{\omega}_2 \in \tilde{\Omega}_2$. Without loss of generality, we can assume that there is a set $\tilde{A}_2 \in \mathcal{A}_2$ such that

$$\tilde{A}_2 \cap \tilde{\Omega}_2 \ = \ \{\tilde{\omega}_2\} \tag{3.22}$$

Otherwise, we can suitably change $\tilde{\Omega}_2$ according to Lemma 8.31. Applying such a set $\tilde{A}_2 \in \mathcal{A}_2$, it follows that

$$\alpha_{\tilde{\omega}_2} \overset{(3.22)}{=} \sum_{\tilde{\omega}_2' \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2'} \delta_{\tilde{\omega}_2'}(\tilde{A}_2) \overset{(3.21)}{=} T(I_{\tilde{A}_2}) \;\in\; \mathcal{L}_\infty(\Omega_1, \mathcal{A}_1)$$

and $\alpha_{\tilde{\omega}_2}(\omega_1) \;=\; T(I_{\tilde{A}_2})(\omega_1) \;\geq\; 0 \quad \forall\, \omega_1 \in \Omega_1\,.$

Finally,

$$\sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2} \;=\; \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2} \delta_{\tilde{\omega}_2}(\Omega_2) \overset{(3.21)}{=} T(I_{\Omega_2}) \;=\; I_{\Omega_1} \;\equiv\; 1$$

□

Le Cam (1986) deals with precise models $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ and $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$. So, it is enough to consider transitions

$$\sigma :\;\; L(\mathcal{E}) \;\rightarrow\; L(\mathcal{F})$$

between the L-spaces $L(\mathcal{E})$ and $L(\mathcal{F})$ generated by $\mathcal{E}$ and $\mathcal{F}$ respectively; cf. Subsection 3.4.1. Accordingly, Le Cam uses the following definition of equivalence:

Definition 3.38 (LC-equivalence) *Precise models $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ and $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$ are called* LC-equivalent*, if there are transition*

$$\sigma_1 \;:\; L(\mathcal{E}) \;\rightarrow\; L(\mathcal{F}) \qquad \textit{and} \qquad \sigma_2 \;:\; L(\mathcal{F}) \;\rightarrow\; L(\mathcal{E})$$

such that

$$\sigma_1(P_\theta) \;=\; Q_\theta \qquad \textit{and} \qquad \sigma_2(Q_\theta) \;=\; P_\theta \qquad\quad \forall\, \theta \in \Theta$$

Confer (Le Cam, 1986, Definition 2.3.1, p. 19 and Theorem 2.3.2).

In contrast to this setup, we have to deal with large sets $(\mathcal{M}_\theta)_{\theta\in\Theta}$ of precise models in the theory of imprecise probabilities. So, it is more convenient to deal with transitions / generalized randomizations

$$\sigma :\;\; \mathrm{ba}(\mathcal{X}, \mathcal{A}) \;\rightarrow\; \mathrm{ba}(\mathcal{Y}, \mathcal{B})$$

between the whole spaces $\mathrm{ba}(\mathcal{X}, \mathcal{A}) \supset L(\mathcal{E})$ and $\mathrm{ba}(\mathcal{Y}, \mathcal{B}) \supset L(\mathcal{F})$. However, this does not change anything. Especially, this has no effect on the definition of equivalence:

Proposition 3.39 *Precise models $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ and $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$ are LC-equivalent if and only if they are equivalent in the sense of Definition 3.15.*

Proof: In the setup of the present book, precise models $\mathcal{E}$ and $\mathcal{F}$ are defined to be (certain) subsets of $\mathrm{ba}(\mathcal{X}, \mathcal{A})$ and $\mathrm{ba}(\mathcal{Y}, \mathcal{B})$ respectively. ($\mathcal{X}$ and $\mathcal{Y}$ are sets with algebras $\mathcal{A}$ and $\mathcal{B}$ respectively.)

Firstly, let $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ and $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$ be LC-equivalent. That is, there are transitions

$$\tilde{\sigma}_1 \;:\; L(\mathcal{E}) \;\rightarrow\; L(\mathcal{F}) \qquad \text{and} \qquad \tilde{\sigma}_2 \;:\; L(\mathcal{F}) \;\rightarrow\; L(\mathcal{E})$$

such that

$$\tilde{\sigma}_1(P_\theta) \;=\; Q_\theta \qquad \text{and} \qquad \tilde{\sigma}_2(Q_\theta) \;=\; P_\theta \qquad\quad \forall\, \theta \in \Theta$$

Note that the L-spaces $L(\mathcal{E})$ and $L(\mathcal{F})$ are bands in $\mathrm{ba}(\mathcal{X},\mathcal{A})$ and $\mathrm{ba}(\mathcal{Y},\mathcal{B})$ respectively. According to Lemma 8.30 a), $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ can be extended to transitions

$$\sigma_1 \;:\; \mathrm{ba}(\mathcal{X},\mathcal{A}) \;\rightarrow\; \mathrm{ba}(\mathcal{Y},\mathcal{B}) \qquad \text{and} \qquad \sigma_2 \;:\; \mathrm{ba}(\mathcal{Y},\mathcal{B}) \;\rightarrow\; \mathrm{ba}(\mathcal{X},\mathcal{A})$$

such that

$$\sigma_1(P_\theta) \;= \tilde{\sigma}_1(P_\theta) \;=\; Q_\theta \qquad \text{and} \qquad \sigma_2(Q_\theta) \;= \tilde{\sigma}_2(Q_\theta) \;=\; P_\theta \qquad \forall\, \theta \in \Theta$$

because $(P_\theta)_{\theta\in\Theta} \subset L(\mathcal{E})$ and $(Q_\theta)_{\theta\in\Theta} \subset L(\mathcal{F})$.

Since the transitions σ_1 and σ_2 are generalized randomizations (cf. Proposition 3.36), it follows that $\mathcal{E}$ and $\mathcal{F}$ are equivalent in the sense of Definition 3.15.

In order to prove the converse statement, let $\mathcal{E} = (P_\theta)_{\theta\in\Theta}$ and $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$ be equivalent in the sense of Definition 3.15.

According to Definition 3.15 and Proposition 3.36, there are transitions

$$\sigma_1 \;:\; \mathrm{ba}(\mathcal{X},\mathcal{A}) \;\rightarrow\; \mathrm{ba}(\mathcal{Y},\mathcal{B}) \qquad \text{and} \qquad \sigma_2 \;:\; \mathrm{ba}(\mathcal{Y},\mathcal{B}) \;\rightarrow\; \mathrm{ba}(\mathcal{X},\mathcal{A})$$

such that

$$\sigma_1(P_\theta) \;=\; Q_\theta \qquad \text{and} \qquad \sigma_2(Q_\theta) \;=\; P_\theta \qquad \forall\, \theta \in \Theta$$

Then, it follows from Lemma 8.30 b) that σ_1 and σ_2 can be restricted to transitions

$$\tilde{\sigma}_1 \;:\; L(\mathcal{E}) \;\rightarrow\; L(\mathcal{F}) \qquad \text{and} \qquad \tilde{\sigma}_2 \;:\; L(\mathcal{F}) \;\rightarrow\; L(\mathcal{E})$$

such that

$$\tilde{\sigma}_1(P_\theta) \;=\; \sigma_1(P_\theta) \;=\; Q_\theta \qquad \text{and} \qquad \tilde{\sigma}_2(Q_\theta) \;=\; \sigma_2(Q_\theta) \;=\; P_\theta \qquad \forall\, \theta \in \Theta$$

That is, $\mathcal{E}$ and $\mathcal{F}$ are LC-equivalent. □

Chapter 4

Least favorable models

4.1 Introduction

4.1.1 Outline

In data-based decision theory, uncertainty is frequently modeled by a prior distribution π and a family of distributions of the observations, $(Q_\theta)_{\theta\in\Theta}$. The prior distribution π describes the uncertainty about the states of natures $\theta \in \Theta$, i.e.: What state of nature θ will be effectively given? With respect to this (unknown) true θ, the distribution of the observations describes the uncertainty about the data which will be observed.
Of course, in practical applications, it is rather unrealistic that precise distributions π and $(Q_\theta)_{\theta\in\Theta}$ are known exactly. Therefore, we allow for a whole set $\mathcal{P}$ of possible precise prior distributions π and whole sets $\mathcal{M}_\theta$ of possible precise distributions Q_θ. According to Section 3.2, we have to search for a randomized decision function τ which minimizes the twofold upper expectation

$$\sup_{\pi\in\mathcal{P}} \int_\Theta \sup_{Q_\theta\in\mathcal{M}_\theta} \int_{\mathcal{Y}} \int_{\mathbb{D}} W_\theta(t)\tau_y(dt)\, Q_\theta(dy)\, \pi(d\theta)$$

Unfortunately, a direct solution of this problem is quite often computationally intractable. However, there are situations where this becomes tractable, namely in the presence of a so-called *least favorable model.*

Most of the research concerning least favorable models was encourage by the celebrated article Huber and Strassen (1973). Huber and Strassen (1973) deals with hypothesis testing where a (rather special) upper prevision is tested against another one. In the context of hypothesis testing, least favorable models are usually called *least favorable pairs.* Testing between coherent upper previsions is equivalent to testing between their credal sets $\mathcal{M}_0$ and $\mathcal{M}_1$. Huber and Strassen (1973) shows that there is a pair $(Q_0, Q_1) \in \mathcal{M}_0 \times \mathcal{M}_1$ which is least favorable: That is, testing between Q_0 and Q_1 is as hard as testing between $\mathcal{M}_0$ and $\mathcal{M}_1$ and, as a consequence, there is an optimal test between Q_0 and Q_1 which is also an optimal test between $\mathcal{M}_0$ and $\mathcal{M}_1$. That way, testing between $\mathcal{M}_0$ and $\mathcal{M}_1$ can be done by testing only between Q_0 and Q_1. This reduces the computational effort substantially. As stated above, it is one of the most important drawbacks of data-based decision theory (including hypothesis testing) under imprecise probabilities that the computational effort of direct solutions is frequently not manageable. Therefore, least favorablility has attracted enormous attention after the publication of Huber and Strassen (1973). For a review of Huber and Strassen (1973) and the work following Huber and

Strassen (1973), confer Augustin (2002). In quite general data-based decision theory, where there are n states of nature (instead of two as in hypothesis testing), an analogous question of that one solved by Huber and Strassen (1973) is: Does there exist a model $(Q_1, Q_2, \dots, Q_n) \in \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_n$ which is simultaneously least favorable for a set of precise prior distributions[1]? This is not always the case but the seminal article Buja (1984) proves a necessary and sufficient condition for the existence of such simultaneously least favorable models in case of upper expectations.
Unfortunately, Buja (1984) contains an error which reduces its applicability significantly. Subsection 4.1.2 highlights the wrong statement, gives a counterexample and discusses the consequences. The validity of the conclusions in Buja (1984) can only be guaranteed by adding a restrictive assumption on the involved upper previsions.

Next, Section 4.2 follows the lines of Buja (1984) - but within the concept of Walley (1991) which dispenses with σ-additivity: While Buja (1984) considers upper expectations only, we use coherent upper previsions. It is shown that the same result as in Buja (1984) is possible without any additional assumption on the involved (coherent) upper previsions. Surprisingly, most of the proofs are similar to those given in Buja (1984). This demonstrates that, in Buja (1984), insistence on σ-additivity of probabilities happens to be an unnecessary burden. By ignoring σ-additivity, we are in line with Le Cam's decision theoretic framework (cf. Le Cam (1964), Le Cam (1986) and Section 3.3) which provides us with the effective methods developed in Section 3.3. Especially, the use of generalized randomizations is crucial.
Subsection 4.2.1 shows how minimal Bayes risks can be calculated and expressed in terms of standard models. Subsection 4.2.2 contains a generalization of the LeCam-Blackwell-Sherman-Stein-Theorem. This theorem plays an important role in Subsection 4.2.3 where the analogue to (Buja, 1984, Theorem 8.2) is proven which characterizes the existence of least favorable models. This is the main theorem of Section 4.2. Subsection 4.2.4 explains how least favorability could be used to deal with situations where the distribution of the data as well as the prior is assumed to be imprecise. These results exemplifies that the classical setup based on Polish spaces and σ-additive probability measures frequently leads to unpleasant – and unnecessary – difficulties.

As a special case, Section 4.3 considers hypothesis testing. Firstly, Susection 4.3.1 explains (and proves) how hypothesis testing fits into the decision theoretic framework. This is not entirely trivial in case of imprecise probabilities. Secondly, the existence of least favorable pairs in case of general coherent upper previsions is proven in Subsection 4.3.3. This result which has already been proven in Baumann (1968) follows as an easy corollary.

As Huber and Strassen (1973), the present chapter is only concerned with the *existence* of least favorable models (and pairs) but an algorithm for explicit calculations has not yet been developed. After Huber and Strassen (1973), a lot of work was done in order to construct least favorable pairs in hypothesis testing (e.g. Rieder (1977), Österreicher (1978), Hafner (1992), Augustin (1998)). In the more general case of the present chapter, this is a matter of further research.

As mentioned above, the following subsection investigates an error in Buja (1984), gives a (counter-)example and shows how the proof of the main theorem in Buja (1984) can be corrected under an additional assumption.

[1] or equivalently: for a class of loss functions; cf. Subsection 4.2.4.

4.1.2 On an error in Buja (1984)

4.1.2.1 Buja's statement and a counter-example

Buja (1984) considers upper expectations on Polish spaces as treated in Subsection 2.4.3.

Let Ξ be a Polish space with Borel-σ-algebra $\mathfrak{B}$. Let $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ be the set of all probability measures on $(\Xi,\mathcal{B})$ and let $\mathcal{C}_\mathrm{b}(\Xi)$ be the Banach space of all bounded continuous function with norm $\|\cdot\|_\infty$.

As in Subsection 2.4.3, $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ is provided with the relative topology on $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ generated by the $\mathcal{C}_\mathrm{b}(\Xi)$ - topology on $\mathrm{ca}(\Xi,\mathfrak{B})$. This topology is usually called *weak topology on* $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$. It is the smallest topology on $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ so that

$$\mathrm{ca}_1^+(\Xi,\mathfrak{B}) \;\to\; \mathbb{R} \qquad P \;\mapsto\; P[f]$$

is continuous for every $f \in \mathcal{C}_\mathrm{b}(\Xi)$. $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ provided with the weak topology is a Polish space. Especially, there is a metric d on $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ which induces the weak topology. Therefore, the weak topology is characterized by sequential convergence. A sequence $(P_n)_{n\in\mathbb{N}}$ in $\mathrm{ca}_1^+(\Xi,\mathfrak{B})$ converges weakly to some $P \in \mathcal{K}(\Xi,\mathcal{B})$ if and only if

$$\lim_n P_n[f] \;=\; p[f] \qquad \forall\, f \in \mathcal{C}_\mathrm{b}(\Xi)$$

The following statement is contained in (Buja, 1984, Proposition 2.1):

> Let $\mathcal{P}$ be the structure of an upper expectation $\overline{P} : \mathcal{L}_\infty(\mathcal{Y},\mathfrak{B}) \to \mathbb{R}$. Assume that $\mathcal{P}$ is tight.
> Then, $\overline{P}[f_n] \searrow \overline{P}[f]$ for every sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}_\mathrm{b}(\Xi)$ such that $f_n \searrow f$ pointwise and $f \in \mathcal{L}_\infty(\mathcal{Y},\mathfrak{B})$.

This is not right as can be seen from the following counterexample:

$\Xi = \mathbb{R}$ is a Polish space with Borel-σ-Algebra $\mathfrak{B} = \mathbb{B}$. Put

$$\mathcal{P} = \Big\{ P \in \mathrm{ca}_1^+(\mathbb{R},\mathbb{B}) \;\Big|\; P\big([0,1)\big) = 1 \Big\} \tag{4.1}$$

Then, define an upper expectation by

$$\overline{P}[f] = \sup_{P\in\mathcal{P}} P[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathbb{R},\mathbb{B})$$

[1] $\mathcal{P}$ is the structure of $\overline{P}$:

For every $P \in \mathrm{ca}_1^+(\mathbb{R},\mathbb{B})$ with $P[f] \le \overline{P}[f]$ $\forall\, f \in \mathcal{L}_\infty(\mathbb{R},\mathbb{B})$, it follows that

$$1 \;=\; -\overline{P}[-I_{[0,1)}] \;\le\; -P[-I_{[0,1)}] \;=\; P\big([0,1)\big) \;\le\; \overline{P}[I_{[0,1)}] \;=\; 1$$

Hence, $P \in \mathcal{P}$.

[2] $\mathcal{P}$ is tight:

For $\varepsilon > 0$ put $K = [0,1]$. Then, K is compact in $\mathbb{R}$ and

$$\sup_{P\in\mathcal{P}} P\big(\mathbb{R}\setminus K\big) \;=\; 1 - \inf_{P\in\mathcal{P}} P(K) \;=\; 1 - 1 \;<\; \varepsilon$$

[3] Put

$$f_n : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^n I_{[0,1]}(x) + I_{(1,\infty)}(x), \qquad n \in \mathbb{N}$$

Then, $(f_n)_{n\in\mathbb{N}} \subset \mathcal{C}_\mathrm{b}(\mathbb{R})$ and

$$\lim_{n\to\infty} f_n(x) = I_{[1,\infty)}(x) \qquad \forall\, x \in \mathbb{R}$$

Because of

$$f_n(x) = x^n \geq x^n \cdot x = f_{n+1}(x) \quad \forall\, x \in [0,1], \qquad \forall\, n \in \mathbb{N}$$

we have $f_n \searrow I_{[1,\infty)} =: f$ pointwise.

[4] $\overline{P}[f_n] = 1 \;\forall\, n \in \mathbb{N}$:

For every $x \in [0,1)$, the Dirac measure δ_x is in $\mathcal{P}$. Hence, $1 \geq f_n$ implies

$$1 \geq \sup_{P\in\mathcal{P}} P[f_n] \geq \lim_{x\nearrow 1} \delta_x[f_n] = \lim_{x\nearrow 1} f_n(x) = \lim_{x\nearrow 1} x^n = 1 \qquad \forall\, n \in \mathbb{N}$$

[5] $\overline{P}[f] = 0$:

$$\overline{P}[f] = \sup_{P\in\mathcal{P}} P[I_{[1,\infty)}] = \sup_{P\in\mathcal{P}} P\big([1,\infty)\big) = 0$$

Accordingly, the following statement which is implicitly contained in (Buja, 1984, Proposition 2.1 and 2.2) does not hold:

> "Every tight structure is weakly compact."

A counterexample is provided by the following:

Choose $\Xi = \mathbb{R}$, $\mathcal{B} = \mathbb{B}$ and $\mathcal{M}_0$ as in (4.1). Then, $\mathcal{P}$ is a tight structure of an upper expectation (cf. [1] and [2]).
Put $P_n := \delta_{1-1/n} \;\forall\, n \in \mathbb{N}$. Then, $(P_n)_{n\in\mathbb{N}} \subset \mathcal{P}$ and $(P_n)_{n\in\mathbb{N}}$ converges weakly to δ_1 in $\mathrm{ca}_1^+(\mathbb{R},\mathbb{B})$ because

$$\lim_n P_n[f] = \lim_n f\left(1 - \frac{1}{n}\right) = f(1) = \delta_1[f] \qquad \forall\, f \in \mathcal{C}_\mathrm{b}(\mathbb{R})$$

However, δ_1 is not an element of $\mathcal{P}$ because $\delta_1\big([0,1)\big) = 0$. Hence, $\mathcal{P}$ is not weakly closed and, therefore, not weakly compact.

The correct statements have already been given by Theorem 2.31 and Theorem 2.32. These theorems show that the reason for the incorrect statement is a permutation of assumption a) in Theorem 2.31 and assumption a) in Theorem 2.32. Indeed, these assumptions seem to be nearly the same. The only difference is that, in Theorem 2.32, it is additionally assumed that the limit f of $(f_n)_{n\in\mathbb{N}}$ is continuous. This difference really matters as can be seen from the above counter-example.

4.1.2.2 Consequences

In Buja (1984), imprecise models are called “approximate models” and denoted by $(\nu_\theta)_{\theta\in\Theta}$. As stated above, each ν_θ is an upper expectation on a Polish space where the corresponding structures $\mathcal{P}_\theta$ have to be compact in the weak topology on ca_1^+. To this end, it is not sufficient to assume tightness of $\mathcal{P}_\theta$ as shown in Subsection 4.1.2.1. Weak closedness of $\mathcal{P}_\theta$ is another necessary additional assumption (cf. also Theorem 2.29). Recall that weak closedness of structures is characterized by Proposition 2.33.

Under this additional assumption, all results of Buja (1984) are valid. However, the proof of the main theorem, (Buja, 1984, Theorem 8.2), has to be revised:

It seems to be not obvious that an approximate model $(\nu_\theta)_{\theta\in\Theta}$ with weakly compact structures $\mathcal{P}_\theta$ induces an upper standard functional whose corresponding structure is also weakly compact. As a consequence, it does not seem to be assured yet if the following is true:

> “it is possible to construct a standard measure S under $s^{(\nu)}$ which equals $s^{(\nu)}$ on the cone $\mathcal{K}$” (Buja, 1984, p. 382).

However, it is possible to give a slightly different proof of (Buja, 1984, Theorem 8.2) so that this problem can be ignored. In the following, the notation is completely adopted from Buja (1984).

Revised proof of (Buja, 1984, Theorem 8.2):

The implication “a) $\Rightarrow$ b)” is not affected. So, it remains to proof the implication “a) $\Leftarrow$ b)”.

Define a linear functional S on the linear space $\mathcal{K}-\mathcal{K}$ by

$$S[k_1-k_2]=s^{(\nu)}[k_1]-s^{(\nu)}[k_2]$$

Additivity of $s^{(\nu)}$ on $\mathcal{K}$ implies that this definition is independent of the special representation k_1-k_2. Subadditivity of $s^{(\nu)}$ on $\mathcal{K}-\mathcal{K}$ implies

$$S[k] \;\le\; s^{(\nu)}[k] \qquad \forall\, k\in\mathcal{K}-\mathcal{K} \tag{4.2}$$

Hahn-Banach (Dunford and Schwartz, 1958, Theorem II.3.10) yields an extension of S on $\mathcal{C}(K)$, so that

$$S[k] \;\le\; s^{(\nu)}[k] \qquad \forall\, k\in\mathcal{C}(K) \tag{4.3}$$

So far, the present proof coincides with that one given in Buja (1984). Since it is not yet assured if $s^{(\nu)}$ is compactly generated, we do not apply (Buja, 1984, Proposition 2.2f)) in the following. (This is the main difference to the proof given in Buja (1984).)

An application of the Riesz representation theorem (Dunford and Schwartz, 1958, Theorem IV.6.3) yields an extension of S on $\mathcal{L}_\infty(K)$ (again denoted by S) which is a probability measure on K.

In contrast to (Buja, 1984, p. 382), we do not state that S would be dominated by $s^{(\nu)}$ on $\mathcal{L}_\infty(K)$. Equation (4.3) is enough to proceed by following the lines of (Buja, 1984, p. 382) again:

(4.3) implies $\int z_\theta \, dS \;=\; \frac{1}{|\Theta|} \;\; \forall \, \theta \in \Theta$. Therefore, S is a standard measure which determines a unique standard model $(S_\theta)_{\theta\in\Theta}$ so that S is the standard measure of $(S_\theta)_{\theta\in\Theta}$ (cf. (Buja, 1984, Section 5)). (Buja, 1984, Theorem 7.1) and (4.3) implies that $(S_\theta)_{\theta\in\Theta}$ is worst-case-sufficient for $(\nu_\theta)_{\theta\in\Theta}$. That is, there is a model $(Q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$ so that $Q_\theta \;\le\; \nu_\theta$ on $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$, $\forall \, \theta \in \Theta$, and so that $(S_\theta)_{\theta\in\Theta}$ is sufficient for $(Q_\theta)_{\theta\in\Theta}$. Hence, for all $k \in \mathcal{K}$,

$$s^{(\nu)}[k] = S[k] \le S^{(Q)}[k] \le s^{(\nu)}[k]$$

where the first inequality follows from (Buja, 1984, Corollary 7.2). That is, $(Q_\theta)_{\theta\in\Theta}$ is least favorable on $\mathcal{K}$. □

Since the results in Buja (1984) are still valid under the additional assumption that the structures have to be weakly closed, the question arises if this additional assumption is restrictive. The simple example in Subsection 4.1.2.1 indicates that, indeed, this is restrictive. This statement is supported by Proposition 2.33 which says that a structure is weakly compact if and only if it can be written as

$$\mathcal{P} \;=\; \left\{ P \in \mathrm{ca}_1^+(\Xi,\mathfrak{B}) \;\middle|\; P[f] \le \overline{P}[f] \;\; \forall \, f \in \mathcal{C}_\mathrm{b}(\Xi) \right\}$$

That is, an upper expectation $\overline{P}$ whose structure is weakly closed is completely defined by its values on $\mathcal{C}_\mathrm{b}(\Xi)$. However, one of the most important special case of upper expectations are F-probabilities which are defined by their values on some $\mathcal{K} \subset \left\{ I_B \;\middle|\; B \in \mathfrak{B} \right\}$. Hence: Whenever we define an upper expectation also by its value on at least one set $B \in \mathfrak{B}$ (i.e. by its value on at least one indicator function), it is not clear if its structure is weakly closed. In consideration of the simple example in Subsection 4.1.2.1, it will usually not be weakly closed. At least, it may be possible to establish easy conditions on $\mathcal{K} \subset \left\{ I_B \;\middle|\; B \in \mathfrak{B} \right\}$ which ensure weak closedness of the structure. Proposition 2.39 shows that the restriction to compact subsets $K = B \subset \Xi$ in $\mathcal{K} \subset \left\{ I_B \;\middle|\; B \in \mathfrak{B} \right\}$ does *not* yield such a sufficient condition.

The following section follows the lines of Buja (1984) and proves that essentially the same result is true if the setup of upper expectations on Polish spaces is replaced by the more general setup of coherent upper previsions on arbitrary sample spaces $(\mathcal{Y},\mathcal{B})$. Essentially, this means that σ-additivity is dropped and that the proofs require some results from Le Cam (1986) as prepared in Section 3.3. Surprisingly, no additional assumption on the involved coherent upper previsions is needed in order to obtain the analogous results as in Buja (1984) even though most of the proofs are quite similar. This shows that insistence on σ-additivity is an unnecessary burden in Buja (1984) and that it can, indeed, be useful to go over to the more general setup of coherent upper prevision.

4.2 Decision problems with n states of nature

4.2.1 Minimal Bayes Risks

4.2.1.1 Introduction

Within the whole Subsection 4.2.1, let $\Theta = \{\theta_1, \ldots, \theta_n\}$ be a finite index set with cardinality n and let π be a precise prior distribution on $(\Theta, 2^\Theta)$, i.e. π is a linear prevision on

$\mathcal{L}_\infty(\Theta, 2^\Theta)$. Put $\pi_\theta := \pi[I_{\{\theta\}}]$.
Let $(\mathbb{D}, \mathcal{D})$ be a decision space and $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ be a loss function

$$\Theta \times \mathbb{D} \;\to\; \mathbb{R}\,, \qquad (\theta, t) \;\mapsto\; W_\theta(t)$$

Finally, let $\mathcal{Y}$ be a set with algebra $\mathcal{B}$. The sample space $(\mathcal{Y}, \mathcal{B})$ represents the possible outcomes of an experiment.

A *decision procedure* is a restricted / ordinary / generalized randomization

$$\sigma : \mathrm{ba}(\mathcal{Y}, \mathcal{B}) \;\to\; \mathrm{ba}(\mathbb{D}, \mathcal{D})$$

– confer Subsection 3.3.1.2.
According to Section 3.2, the Bayes risk of an ordinary randomization $\sigma \in \mathcal{T}_0(\mathcal{Y}, \mathbb{D})$ defined by a finitely additive Markov kernel

$$\mathcal{Y} \times \mathcal{D} \;\to\; \mathbb{R}\,, \qquad (y, D) \;\mapsto\; \tau_y(D)$$

via (3.2) is equal to

$$\begin{aligned} R_\pi\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big) \;&=\; \int_{\theta\in\Theta} \sup_{Q_\theta \in \mathcal{M}_\theta} \int_{\mathcal{Y}} \int_{\mathbb{D}} W_\theta(t)\,\tau_y(dt)\,Q_\theta(dy)\,\pi(d\theta) \;=\\ &=\; \sum_{\theta\in\Theta} \pi_\theta \cdot \sup_{Q_\theta\in\mathcal{M}_\theta} \sigma(Q_\theta)(W_\theta) \end{aligned}$$

where $(\overline{Q}_\theta)_{\theta\in\Theta}$ is an imprecise model on $(\mathcal{Y}, \mathcal{B})$ and $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of credal sets. Analogously, the Bayes risk can be defined for every generalized randomization $\sigma \in \mathcal{T}(\mathcal{Y}, \mathbb{D})$ as

$$R_\pi\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big) \;=\; \sum_{\theta\in\Theta} \pi_\theta \cdot \sup_{Q_\theta\in\mathcal{M}_\theta} \sigma(Q_\theta)(W_\theta) \tag{4.4}$$

– confer Section 3.3.
Therefore, the minimal Bayes risk

$$\inf_\sigma R_\pi\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big) \tag{4.5}$$

can be calculated over all restricted, ordinary or generalized randomizations σ. It is shown in the following two subsections that the above infimum is equal for all of the three types of randomizations. That is, it does not matter if we also allow for generalized randomizations. Subsection 4.2.1.2 is concerned with precise models $(Q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y}, \mathcal{B})$, Subsection 4.2.1.3 is concerned with imprecise models $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y}, \mathcal{B})$.
Furthermore, it is shown that the term "minimal Bayes risk" is justified because the infimum in (4.5) is attained in a generalized ranomization.
The main goal of the present subsection is to express minimal Bayes risks in terms of standard measures for precise models $(Q_\theta)_{\theta\in\Theta}$ (Theorem 4.4) and in terms of standard upper previsions for imprecise models $(Q_\theta)_{\theta\in\Theta}$ (Theorem 4.7).
Since π is a fixed precise prior distribution in the present subsection, the index is dropped in R_π and the Bayes risk with respect to the fixed π is denoted by $R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big)$.

4.2.1.2 Precise Models

Let $(Q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y},\mathcal{B})$. According to (4.4), the Bayes risk of a generalized randomization $\sigma \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ is

$$R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big) \;=\; \sum_{\theta\in\Theta} \pi_\theta \cdot \sigma(Q_\theta)(W_\theta) \tag{4.6}$$

then. Of course, (4.6) coincides with the usual definition of the Bayes risk if σ is defined by a randomized decision function as in (3.2).
The minimal Bayes risk is the same if we let σ vary among the restricted, ordinary or generalized randomizations:

Proposition 4.1 *There is a generalized randomization $\tilde{\sigma} \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ such that*

$$R\big((Q_\theta)_{\theta\in\Theta},\tilde{\sigma},W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big)$$

Furthermore,

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big)$$

coincides for $\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \;=\; \mathcal{T}_r(\mathcal{Y},\mathbb{D})\,,\; =\; \mathcal{T}_0(\mathcal{Y},\mathbb{D})$ and $=\; \mathcal{T}(\mathcal{Y},\mathbb{D})\,.$

Proof: The definition of the topology of pointwise convergence implies continuity of the map

$$\sigma \;\mapsto\; \big(\sigma(Q_{\theta_1})[W_{\theta_1}]\,,\ldots,\,\sigma(Q_{\theta_n})[W_{\theta_n}]\big)$$

and, therefore, continuity of $\sigma \;\mapsto\; R((Q_\theta)_{\theta\in\Theta},\sigma,W)\,.$
According to (Denkowski et al., 2003, Theorem 1.3.11), continuity of this function implies that it attains its minimum in a generalized randomization $\tilde{\sigma} \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ because $\mathcal{T}(\mathcal{Y},\mathbb{D})$ is compact (Theorem 3.9).

The second statement also follows from continuity of $\sigma \;\mapsto\; R((Q_\theta)_{\theta\in\Theta},\sigma,W)$ because $\mathcal{T}_r(\mathcal{Y},\mathbb{D})$ and $\mathcal{T}_0(\mathcal{Y},\mathbb{D})$ are dense in $\mathcal{T}(\mathcal{Y},\mathbb{D})$ (Theorem 3.10). □

Notation 4.2 *As in Proposition 4.1, it often does not matter, if we consider $\mathcal{T}_r(\mathcal{Y},\mathbb{D})$, $\mathcal{T}_0(\mathcal{Y},\mathbb{D})$ or $\mathcal{T}(\mathcal{Y},\mathbb{D})$. These cases are indicated by the use of the symbol $\mathcal{T}_*(\mathcal{Y},\mathbb{D})$. That is*

$$\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \;\in\; \big\{\mathcal{T}_r(\mathcal{Y},\mathbb{D})\,,\; \mathcal{T}_0(\mathcal{Y},\mathbb{D})\,,\; \mathcal{T}(\mathcal{Y},\mathbb{D})\big\}$$

The following lemma provides an example for the fact that sufficiency is strongly connected with the decision theoretic Bayes risk. In Subsection 4.2.2, this is strengthened in case of imprecise probabilities.

Lemma 4.3 *If a precise model $(P_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A})$ is sufficient for the precise model $(Q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$, then*

$$\inf_{\rho\in\mathcal{T}_*(\mathcal{X},\mathbb{D})} R\big((P_\theta)_{\theta\in\Theta},\rho,W\big) \;\le\; \inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big)$$

Proof: There is some $\psi \in \mathcal{T}(\mathcal{X},\mathcal{Y})$ so that $\psi(P_\theta) = Q_\theta \ \ \forall\, \theta \in \Theta$. Therefore,

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \sigma(Q_\theta)[W_\theta] \ = \ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \sigma\big(\psi(P_\theta)\big)[W_\theta] = $$
$$= \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \big(\sigma\circ\psi\big)(P_\theta)[W_\theta] \ \geq \ \inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \rho(P_\theta)[W_\theta]$$

because $\sigma\circ\psi \ \in \ \mathcal{T}(\mathcal{X},\mathbb{D}) \ \ \forall\,\sigma \in \mathcal{T}(\mathcal{X},\mathbb{D})$. Finally, Proposition 4.1 implies that the inequality is valid for any choice of $\mathcal{T}_*(\mathcal{Y},\mathbb{D})$ and $\mathcal{T}_*(\mathcal{X},\mathcal{Y})$. □

For the loss function

$$W \ : \ \Theta\times\mathbb{D} \ \rightarrow \ \mathbb{R} \qquad (\theta,t) \ \mapsto \ W_\theta(t)$$

put

$$K(W) : \ u \ \mapsto \ \inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\iota_\theta(u) \tag{4.7}$$

on $\mathbb{R}^n$ where $\iota_\theta(u) = u_\theta$ is the θ-component of $u \in \mathbb{R}^\Theta \cong \mathbb{R}^n$. Note that $K(W)$ is concave and, therefore, continuous on $\mathbb{R}^n$. Hence, the restriction of $K(W)$ on $\mathcal{U}$ is Borel-measurable and bounded. Therefore, $S^{\mathcal{F}}\big[K(W)\big]$ is defined well for the standard measure $S^{\mathcal{F}}$ of any precise model $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$.

Theorem 4.4 *Let $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y},\mathcal{B})$ and $S^{\mathcal{F}}$ its standard measure. Then,*

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big) \ = \ S^{\mathcal{F}}\big[K(W)\big]$$

Proof: According to Theorem 3.30, the standard model $\big(S^{\mathcal{F}}_\theta\big)_{\theta\in\Theta}$ is equivalent to $\mathcal{F} = (Q_\theta)_{\theta\in\Theta}$. That is $\big(S^{\mathcal{F}}_\theta\big)_{\theta\in\Theta}$ and $\mathcal{F}$ are mutual sufficient. So, a twofold application of Lemma 4.3 yields

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big(\mathcal{F},\sigma,W\big) \ = \ \inf_{\rho\in\mathcal{T}_*(\mathcal{U},\mathbb{D})} R\big((S^{\mathcal{F}}_\theta)_{\theta\in\Theta},\rho,W\big)$$

and an application of Lemma 8.32 closes the proof. □

4.2.1.3 Imprecise Models

Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y},\mathcal{B})$ with corresponding structures $\mathcal{M}_\theta$, $\theta\in\Theta$. Let $\overline{S}$ be the standard upper prevision of $(\overline{Q}_\theta)_{\theta\in\Theta}$. According to (4.4), the Bayes risk of a generalized randomization $\sigma \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ is

$$R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big) \ = \ \sum_{\theta\in\Theta} \pi_\theta \cdot \sup_{Q_\theta\in\mathcal{M}_\theta} \sigma(Q_\theta)(W_\theta) \tag{4.8}$$

then. Hence,

$$R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big) \ = \sup_{(Q_\theta)_{\theta\in\Theta}\in(\mathcal{M}_\theta)_{\theta\in\Theta}} R\big((Q_\theta)_{\theta\in\Theta},\sigma,W\big) \tag{4.9}$$

These definitions include that we have chosen the Γ-minimax optimality criterion which represents a worst case consideration (cf. Section 3.1) – as done in Huber and Strassen (1973) and Buja (1984).

Now, we can derive the analogs of Proposition 4.1 and Theorem 4.4 in case of imprecise models:

Proposition 4.5 *There is a generalized randomization* $\tilde{\sigma} \in \mathcal{T}(\mathcal{Y}, \mathbb{D})$ *such that*

$$R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \tilde{\sigma}, W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big)$$

Furthermore,

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big)$$

coincides for $\mathcal{T}_*(\mathcal{Y}, \mathbb{D}) \,=\, \mathcal{T}_r(\mathcal{Y}, \mathbb{D})\,,\; =\, \mathcal{T}_0(\mathcal{Y}, \mathbb{D})$ *and* $=\, \mathcal{T}(\mathcal{Y}, \mathbb{D})\,.$

Actually, we want to deal with decision problems where the prior distribution and the distribution of the observation may be imprecise. This is done in Subsection 4.2.3 whereas, in the present preparatory subsection, the prior distribution is assumed to be precise. In case of imprecise prior distributions, Proposition 4.5 is not enough to keep the connection between ordinary and generalized randomizations. However, it is possible to extend this proposition to imprecise prior distributions in Subsection 4.2.3 (Proposition 4.18).
In order to proof Proposition 4.5, the following important lemma is needed. Essentially, this is an application of the minimax theorem. For this application, compactness of the structures $\mathcal{M}_\theta$ is crucial.

Lemma 4.6

$$\textbf{(a)}\quad \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;=\; \sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \; \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta, \sigma, W\big)$$

$$\textbf{(b)}\quad \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;=\; \sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \; \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta, \sigma, W\big)$$

$$\textbf{(c)}\quad \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;=\; \sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta, \sigma, W\big)$$

Proof:

(a) According to Theorem 2.16, every credal set $\mathcal{M}_\theta$ is compact in the $\mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$-topology on $\mathrm{ba}(\mathcal{Y}, \mathcal{B})$. Then, (Dunford and Schwartz, 1958, Lemma V.3.3, Lemma I.8.2 and Theorem I.8.5) imply that $\prod_{\theta\in\Theta} \mathcal{M}_\theta$ is a compact Hausdorff space[2]. For every $\sigma \in \mathcal{T}_r(\mathcal{Y}, \mathbb{D})$ there is some $\kappa : \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B}) \to \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ so that $\sigma(\mu)[g] = \mu[\kappa(g)]$ for every $g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$, $\mu \in \mathrm{ba}(\mathcal{Y}, \mathcal{B})$; confer (3.1) and Definition 3.7. Hence,

$$\mathcal{M}_\theta \;\to\; \mathbb{R}\,, \qquad Q_\theta \;\mapsto\; \sigma(Q_\theta)[W_\theta]$$

is continuous for every $\theta \in \Theta$ and this implies continuity of the map

$$(Q_\theta)_{\theta\in\Theta} \;\mapsto\; -\sum_{\theta\in\Theta} \pi_\theta \sigma(Q_\theta)[W_\theta] =: \Gamma\big((Q_\theta)_{\theta\in\Theta}, \sigma\big)$$

on $\prod_{\theta\in\Theta} \mathcal{M}_\theta$ for every $\sigma \in \mathcal{T}_r(\mathcal{Y}, \mathbb{D})$. $(Q_\theta)_\theta \mapsto \Gamma\big((Q_\theta)_\theta, \sigma\big)$ is convex on $\prod_{\theta\in\Theta} \mathcal{M}_\theta$ for every $\sigma \in \mathcal{T}_r(\mathcal{Y}, \mathbb{D})$ and $\sigma \mapsto \Gamma\big((Q_\theta)_\theta, \sigma\big)$ is concave on $\mathcal{T}_r(\mathcal{Y}, \mathbb{D})$ for every $(Q_\theta)_{\theta\in\Theta} \in \prod_{\theta\in\Theta} \mathcal{M}_\theta$. Then, the minimax theorem (Fan, 1953, Theorem 2) yields

$$\begin{aligned}
\inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big) \;&=\; -\sup_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} \; \inf_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \Gamma\big((Q_\theta)_\theta, \sigma\big) = \\
&= -\inf_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \; \sup_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} \Gamma\big((Q_\theta)_\theta, \sigma\big) = \\
&= \sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} \; \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta, \sigma, W\big)
\end{aligned}$$

[2] in the n-fold product topology of the $\mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$-topology

(b) and **(c)**: Proposition 4.1 and part (a) of the present lemma yield

$$\begin{aligned}\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big) \;\geq\; &\sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta}\;\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta,\sigma,W\big) = \\ = &\sup_{(Q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta}\;\inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta,\sigma,W\big) = \\ \overset{(a)}{=} &\inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \geq \inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,W\big)\end{aligned}$$

□

Proof of Proposition 4.5: As already stated in the proof of Proposition 4.1, the map $\sigma \mapsto R((Q_\theta)_{\theta\in\Theta},\sigma,W)$ is continuous with respect to the topology of pointwise convergence on $\mathcal{T}(\mathcal{Y},\mathbb{D})$ for every precise model $(Q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$. Hence, it follows from (4.9) that $\sigma \mapsto R((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W)$ is lower semicontinuous. According to (Denkowski et al., 2003, Theorem 1.3.11), lower semicontinuity of this function implies that it attains its minimum in a generalized randomization $\tilde{\sigma}\in\mathcal{T}(\mathcal{Y},\mathbb{D})$ because $\mathcal{T}(\mathcal{Y},\mathbb{D})$ is compact (Theorem 3.9).

The second statement is a direct consequence of Lemma 4.6 (a), Proposition 4.1 and Lemma 4.6 (c). □

Theorem 4.7 *Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y},\mathcal{B})$ and $\overline{S}$ its standard upper prevision. Then,*

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big) \;=\; \overline{S}\Big[K(W)\Big]$$

Proof: This is a direct consequence of Lemma 4.6, Theorem 4.4 and the definition of the standard upper prevision. □

4.2.2 A general LeCam-Blackwell-Sherman-Stein-Theorem

This subsection contains a generalization of the LeCam-Blackwell-Sherman-Stein-Theorem. On the one hand, it is needed in the proof of the main theorem of the present section, Theorem 4.12, on the other hand, it is also interesting of its own because it is a generalization of a family of well-known theorems which were developed during several decades; cf. e.g. Blackwell (1953), Le Cam (1964), Heyer (1969) and Buja (1984).

Let Θ be a finite index set. Let π be a prior distribution on $(\Theta,2^\Theta)$ so that

$$\pi_\theta := \pi[I_{\{\theta\}}] \;>\; 0 \qquad \forall\,\theta\in\Theta \tag{4.10}$$

Let $(P_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{X},\mathcal{A})$ and $(\overline{Q}_\theta)_{\theta\in\Theta}$ an imprecise model on $(\mathcal{Y},\mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of credal sets. Let $S^{(P_\theta)_\theta}$ be the standard measure of $(P_\theta)_{\theta\in\Theta}$ and $\overline{S}$ the standard upper prevision of $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{U},\mathcal{C})$.

Let Ψ be the set of all functions $k\in\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ such that there is some decision space $(\mathbb{D},\mathcal{D})$ and a loss function $W:(\theta,t)\mapsto W_\theta(t)$, $(W_\theta)_{\theta\in\Theta}\subset\mathcal{L}_\infty(\mathbb{D},\mathcal{D})$ where $k(u)=\inf_{\tau\in\mathbb{D}}\sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\iota_\theta(u)\;\;\forall\,u\in\mathcal{U}$.

Since π is again a fixed precise prior distribution in the present subsection, the index is dropped in R_π and the Bayes risk with respect to the fixed π is denoted by $R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big)$.

Theorem 4.8 is the analog to (Buja, 1984, Theorem 7.1) and can also be proven analogously.

Theorem 4.8 *With the above settings and under asumption (4.10), the following statements are equivalent:*

(a) $(P_\theta)_{\theta\in\Theta}$ *is worst-case-sufficient for* $(\overline{Q}_\theta)_{\theta\in\Theta}$.

(b) $S^{(P_\theta)_\theta}[k] \;\leq\; \overline{S}[k] \qquad \forall\, k \in \Psi$

(c) *For every finite decision space* $(\mathbb{D}, \mathcal{D})$ *and every bounded loss function* W*,*

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((P_\theta)_{\theta\in\Theta}, \rho, W\big) \;\leq\; \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big)$$

(d) *For every decision space* $(\mathbb{D}, \mathcal{D})$ *and every bounded loss function* W*,*

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((P_\theta)_{\theta\in\Theta}, \rho, W\big) \;\leq\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta}, \sigma, W\big)$$

Proof: The proof has the following structure: (a)⇔(d), (d)⇔(c), (d)⇔(b)

(a)⇒(d): This is a direct consequence of Lemma 4.3.

(a)⇐(d): Put $\mathbb{D} = \mathcal{Y}$ and $\psi_0(\mu) = \mu \quad \forall\, \mu \in \mathrm{ba}(\mathcal{Y}, \mathcal{B})$. Then (d) implies that for all $(g_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$,

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathcal{Y})} R\big((P_\theta)_\theta, \rho, (g_\theta)_\theta\big) \leq R\big((\overline{Q}_\theta)_\theta, \psi_0, (g_\theta)_\theta\big)$$

which may be rewritten as $\displaystyle\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathcal{Y})} \sum_{\theta\in\Theta} \pi_\theta\Big(\rho(P_\theta)[g_\theta] - \overline{Q}_\theta[g_\theta]\Big) \leq 0\,.$

Put $\Gamma\big(\rho, (g_\theta)_\theta\big) := \sum_{\theta\in\Theta} \pi_\theta\big(\rho(P_\theta)[g_\theta] - \overline{Q}_\theta[g_\theta]\big)$. Then,

$$\sup_{(g_\theta)_\theta\subset\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})} \; \inf_{\rho\in\mathcal{T}(\mathcal{X},\mathcal{Y})} \Gamma\big(\rho, (g_\theta)_\theta\big) \;\leq\; 0 \tag{4.11}$$

$\mathcal{T}(\mathcal{X}, \mathcal{Y})$ is compact, $\rho \mapsto \Gamma\big(\rho, (g_\theta)_\theta\big)$ is continuous and convex, $(g_\theta)_{\theta\in\Theta} \mapsto \Gamma\big(\rho, (g_\theta)_\theta\big)$ is concave. So, the minimax theorem (Fan, 1953, Theorem 2) and (4.11) yield

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathcal{Y})} \; \sup_{(g_\theta)_\theta\subset\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})} \Gamma\big(\rho, (g_\theta)_\theta\big) \;\leq\; 0$$

Compactness of $\mathcal{T}(\mathcal{X}, \mathcal{Y})$ and lower semicontinuity of

$$\rho \;\;\mapsto\;\; \sup_{(g_\theta)_\theta\subset\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})} \Gamma\big(\rho, (g_\theta)_\theta\big)$$

imply the existence of some $\rho_0 \in \mathcal{T}(\mathcal{X}, \mathcal{Y})$ so that

$$\sup_{(g_\theta)_\theta\subset\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})} \Gamma\big(\rho_0, (g_\theta)_\theta\big) \;\leq\; 0 \tag{4.12}$$

(cf. (Denkowski et al., 2003, Theorem 1.3.11)). Since $\pi_\theta > 0 \;\; \forall\, \theta \in \Theta$, it follows from (4.12) that

$$\rho_0(P_\theta)[g_\theta] \;\leq\; \overline{Q}_\theta[g_\theta] \qquad \forall\, g_\theta \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B}) \qquad \forall\, \theta \in \Theta$$

(d)⇒(c): This is obvious.

(d)⇐(c): Let $\sigma : \mu \mapsto \kappa^*(\mu)$ be a restricted randomization from $\mathcal{Y}$ to $\mathbb{D}$ where

$$\kappa^*(\mu)[g] = \mu\Big[\sum_{t\in D} g(t)\alpha_t\Big]$$

and D is a finite subset of $\mathbb{D}$. $(D, 2^D)$ may be regarded as a finite decision space and σ may be regarded as an element of $\mathcal{T}(\mathcal{Y}, D)$. Then, (c) implies

$$\inf_{\hat\rho\in\mathcal{T}(\mathcal{X},D)} R\big((P_\theta)_\theta, \hat\rho, W\big) \;\le\; R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \tag{4.13}$$

Note that this is true for every $\sigma \in \mathcal{T}_r(\mathcal{Y}, D)$. Since every element of $\mathcal{T}_r(\mathcal{X}, D)$ may be regarded as an element of $\mathcal{T}_r(\mathcal{X}, \mathbb{D})$, Proposition 4.1 implies

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((P_\theta)_\theta, \rho, W\big) \;\le\; \inf_{\hat\rho\in\mathcal{T}(\mathcal{X},D)} R\big((P_\theta)_\theta, \hat\rho, W\big) \tag{4.14}$$

Hence, (according to Proposition 4.5)

$$\begin{aligned}\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((P_\theta)_\theta, \rho, W\big) \;&\overset{(4.14),(4.13)}{\le}\; \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;=\\ &=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big)\end{aligned}$$

(d)⇔(b): This is a direct consequence of Theorem 4.4 and Theorem 4.7. □

4.2.3 Least favorable models for n states of nature

Let again the index set Θ be finite with cardinality n. Let π be a prior distribution on $(\Theta, 2^\Theta)$ so that $\pi_\theta := \pi[I_{\{\theta\}}] > 0 \;\; \forall\, \theta \in \Theta$. Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y}, \mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of structures. Let $(\mathbb{D}, \mathcal{D})$ be a fixed decision space and let $\mathcal{W}$ be a set of bounded loss functions

$$W \;:\; (\theta, t) \;\mapsto\; W_\theta(t)\,, \qquad (W_\theta)_{\theta\in\Theta} \;\subset\; \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$$

Definition 4.9 *A model* $(Q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *is called* least favorable (precise) model of $(\mathcal{M}_\theta)_{\theta\in\Theta}$ for $\mathcal{W}$ *if*

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((Q_\theta)_\theta, \sigma, W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \qquad \forall\, W \in \mathcal{W}$$

That is, the minimal Bayes risk of the imprecise model is attained in the least favorable model which represents the worst-case. (This justifies the term "least favorable".) Remember that our definition of the Bayes risk corresponds to a worst-case consideration. We are not primarily interested in a set of loss functions but in a set of prior distributions. However, a set of prior distributions can always be transformed into a set of loss functions (cf. Subsection 4.2.4).
For precise models $\mathcal{F} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$, put

$$\Phi_{\mathcal{F}} := \big\{h \in \mathcal{L}_\infty(\mathcal{U}, \mathcal{C}) \;\big|\; S^{\mathcal{F}}[h] = \overline{S}[h]\big\}$$

where $S^{\mathcal{F}}$ is the standard measure of $\mathcal{F}$ and $\overline{S}$ is the standard upper prevision of $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{U}, \mathcal{C})$.

Lemma 4.10 *$\Phi_{\mathcal{F}}$ is a norm-closed convex cone in $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$.*

Proof: For $h \in \Phi_{\mathcal{F}}$ and $c \in [0,\infty)$, $\overline{S}[ch] = c\overline{S}[h] = cS^{\mathcal{F}}[h] = S^{\mathcal{F}}[ch]\,.$

For $h_1, h_2 \in \Phi_{\mathcal{F}}$,

$$\overline{S}[h_1+h_2] \;\le\; \overline{S}[h_1]+\overline{S}[h_2] \;=\; S^{\mathcal{F}}[h_1]+S^{\mathcal{F}}[h_2] \;=\; S^{\mathcal{F}}[h_1+h_2] \;\le\; \overline{S}[h_1+h_2]$$

For $(h_m)_{m\in\mathbb{N}} \subset \Phi_{\mathcal{F}}$, $\lim_m \|h_m - h\| = 0$ and $h \in \mathcal{L}_\infty(\mathcal{U},\mathcal{C})$,

$$\overline{S}[h] \;\le\; \limsup_m \big(\,\overline{S}[h_m]+\overline{S}[h-h_m]\,\big) \;=\; \limsup_m S^{\mathcal{F}}[h_m] \;=\; S^{\mathcal{F}}[h]$$

i.e. $S^{\mathcal{F}}[h] = \overline{S}[h]$. □

For every decision function

$$W \;:\; (\theta,t) \;\mapsto\; W_\theta(t)\,, \qquad (W_\theta)_{\theta\in\Theta} \;\subset\; \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$$

define $K(W)$ as in (4.7):

$$K(W) = \inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\,\iota_\theta$$

Put

$$\Psi_{\mathcal{W}} := \big\{K(W)\;\big|\;(W_\theta)_\theta \in \mathcal{W}\big\} \;\subset\; \mathcal{L}_\infty(\mathcal{U},\mathcal{C})$$

$\tilde{\Psi}_{\mathcal{W}}$ denotes the smallest norm-closed convex cone in $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ which contains $\Psi_{\mathcal{W}}$. Lemma 4.11 is a direct consequence of Theorem 4.4 and Theorem 4.7:

Lemma 4.11 *$\mathcal{F} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ is least favorable for $\mathcal{W}$ if and only if*

$$S^{\mathcal{F}}[k] \;=\; \overline{S}[k] \qquad \forall\, k \in \Psi_{\mathcal{W}}$$

Theorem 4.12 is the analog to (Buja, 1984, Theorem 8.2). It characterizes the existence of least favorable models in full generality.

Theorem 4.12 *With the above settings and under asumption (4.10), the following statements are equivalent:*

(a) *There is some $\mathcal{F} := (Q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ which is least favorable for $\mathcal{W}$.*

(b) $\overline{S}[k_1+k_2] = \overline{S}[k_1]+\overline{S}[k_2] \qquad \forall\, k_1,k_2 \in \tilde{\Psi}_{\mathcal{W}}$

Proof:

(a)⇒(b): Statement (a) and Lemma 4.11 imply $\Psi_{\mathcal{W}} \subset \Phi_{\mathcal{F}}$. According to Lemma 4.10, $\tilde{\Psi}_{\mathcal{W}} \subset \Phi_{\mathcal{F}}$ and $k_1+k_2 \in \Phi_{\mathcal{F}}\;\; \forall\, k_1,k_2 \in \tilde{\Psi}_{\mathcal{W}}$. Hence, for every $k_1,k_2 \in \tilde{\Psi}_{\mathcal{W}}$

$$\overline{S}[k_1+k_2] \;=\; S^{\mathcal{F}}[k_1+k_2] \;=\; S^{\mathcal{F}}[k_1]+S^{\mathcal{F}}[k_2] \;=\; \overline{S}[k_1]+\overline{S}[k_2]$$

(a)⇐(b): Put $S[k] := \overline{S}[k]\;\; \forall\, k \in \tilde{\Psi}_{\mathcal{W}}$ and

$$S[k_1-k_2] := S[k_1]-S[k_2] = \overline{S}[k_1]-\overline{S}[k_2]$$

for all $k_1, k_2 \in \tilde{\Psi}_{\mathcal{W}}$. Statement (b) implies that this is defined well. Hence, S is a linear functional on the vector space $\mathrm{lin}(\tilde{\Psi}_{\mathcal{W}}) = \tilde{\Psi}_{\mathcal{W}} - \tilde{\Psi}_{\mathcal{W}}$. For every $k = k_1 - k_2 \in \tilde{\Psi}_{\mathcal{W}} - \tilde{\Psi}_{\mathcal{W}} = \mathrm{lin}(\tilde{\Psi}_{\mathcal{W}})$,

$$S[k] \;=\; \overline{S}[k_2 + k_1 - k_2] - \overline{S}[k_2] \;\le\; \overline{S}[k_2] + \overline{S}[k_1 - k_2] - \overline{S}[k_2] \;=\; \overline{S}[k]$$

Due to the Hahn-Banach-Theorem ((Dunford and Schwartz, 1958, Theorem II.3.10)), S can be extended to a linear functional on $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ (again denoted by S) so that

$$S[h] \le \overline{S}[h] \qquad \forall\, h \in \mathcal{L}_\infty(\mathcal{U},\mathcal{C}) \tag{4.15}$$

(4.15) implies, that $S\big[I_{\mathcal{U}}\big] = 1$ and $S[\iota_\theta] = \frac{1}{n} \quad \forall\,\theta \in \Theta$ (cf. Theorem 3.30). Then, $S_\theta : h \mapsto S[n\iota_\theta h]$ defines a precise model $(S_\theta)_{\theta\in\Theta}$ on $(\mathcal{U},\mathcal{C})$. For every decision space $(\hat{\mathbb{D}}, \hat{\mathcal{D}})$ and every $\hat{W} : (\theta, t) \mapsto \hat{W}_\theta(t)\,,\ (\hat{W}_\theta)_\theta \subset \mathcal{L}_\infty(\hat{\mathbb{D}}, \hat{\mathcal{D}})$,

$$\inf_{\rho\in\mathcal{T}(\mathcal{U},\hat{\mathbb{D}})} R\big((S_\theta)_\theta, \rho, \hat{W}\big) = S\big[K\big(\hat{W}\big)\big] \tag{4.16}$$

according to Lemma 8.32 and

$$\begin{aligned}\inf_{\rho\in\mathcal{T}(\mathcal{U},\hat{\mathbb{D}})} R\big((S_\theta)_\theta, \rho, \hat{W}\big) \;&\overset{(4.16)}{=}\; S\big[K\big(\hat{W}\big)\big] \;\overset{(4.15)}{\le}\; \overline{S}\big[K\big(\hat{W}\big)\big] \;= \\ &=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\hat{\mathbb{D}})} R\big((\overline{Q}_\theta)_\theta, \sigma, \hat{W}\big)\end{aligned}$$

according to Theorem 4.7. Hence, Theorem 4.8 implies that $(S_\theta)_{\theta\in\Theta}$ is worst-case-sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$, i.e. there is some $\rho \in \mathcal{T}(\mathcal{U},\mathcal{Y})$ so that $Q_\theta := \rho(S_\theta) \in \mathcal{M}_\theta \ \ \forall\,\theta \in \Theta$. Finally for all $W \in \mathcal{W}$,

$$\begin{aligned}\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;&=\; \overline{S}\Big[K(W)\Big] \;=\; S\Big[K(W)\Big] \;= \\ \overset{(4.16)}{=}\;\; &\inf_{\rho\in\mathcal{T}(\mathcal{U},\mathbb{D})} R\big((S_\theta)_\theta, \rho, W\big) \;\le\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((Q_\theta)_\theta, \sigma, W\big)\end{aligned}$$

where the last inequality follows from Lemma 4.3. □

4.2.4 Application of Least Favorable Models

Situations where we are faced with one precise prior distribution and a set of loss functions seem to be of secondary interest. More frequently, we are interested in situations where we are faced with an *imprecise* prior and one fixed loss function. However, the second issue can be treated as a special case of the first one:
Let Θ be a finite index set with cardinality n and

$$W:\ (\theta, t) \;\mapsto\; W_\theta(t)\,, \qquad (W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$$

be a loss function. Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y},\mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of structures. Let $\overline{\Pi}$ be a coherent upper prevision on $\mathcal{L}_\infty(\Theta, 2^\Theta)$ i.e. $\overline{\Pi}$ corresponds to a set of prior distributions $\mathcal{P} := \big\{\pi \in \mathrm{ba}(\Theta, 2^\Theta)\,\big|\, \pi[a] \le \overline{\Pi}[a]\ \forall\, a \in \mathcal{L}_\infty(\Theta, 2^\Theta)\big\}$.

For some $\pi \in \mathcal{P}$, put $\pi_\theta := \pi[I_{\{\theta\}}] \;\; \forall\, \theta \in \Theta$. Let σ be a (generalized) randomization. For the prior π, the Bayes risk is

$$\begin{aligned} R_\pi\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \;&=\; \sum_{\theta\in\Theta} \pi_\theta \sigma(\overline{Q}_\theta)[W_\theta] \;=\; \frac{1}{n}\sum_{\theta\in\Theta} \sigma(\overline{Q}_\theta)[n\pi_\theta W_\theta] \;= \\ &=\; R_0\big((\overline{Q}_\theta)_\theta, \sigma, W^{(\pi)}\big) \end{aligned}$$

where $R_0\big((\overline{Q}_\theta)_\theta, \sigma, (n\pi_\theta W_\theta)_\theta\big)$ denotes the Bayes risk for the uniform prior π_0 defined by $\pi_0[I_\theta] = \frac{1}{n}$ and $W^{(\pi)}$ denotes the loss function

$$W^{(\pi)}: \;\; (\theta, t) \;\mapsto\; n\pi_\theta W_\theta(t)$$

That is every prior can be absorbed in the loss function. So, we can transform the set $\mathcal{P}$ of priors π into a set $\mathcal{W}$ of loss functions $(n\pi_\theta W_\theta)_{\theta\in\Theta}$. Next, Theorem 4.12 yields a necessary and sufficient condition for the existence of a precise model which is simultaneously least favorable for the set of loss functions $\mathcal{W}$. We may also say that such a precise model is *simultaneously least favorable for the set of priors* $\mathcal{P}$. Note, that assumption (4.10) is fulfilled for the prior $\pi_0[I_\theta] = \frac{1}{n}$ so that Theorem 4.12 is always applicable for any set of priors $\mathcal{P}$.

Definition 4.13 *A model* $(Q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *is called* least favorable (precise) model of $(\mathcal{M}_\theta)_{\theta\in\Theta}$ for the set of priors $\mathcal{P}$ *if*

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((Q_\theta)_\theta, \sigma, W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \qquad \forall\, \pi \in \mathcal{P}$$

According to the preceding paragraphs, we have the following proposition:

Proposition 4.14 $(Q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *is a least favorable (precise) model of* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *for the set of priors* $\mathcal{P}$ *if and only if it is a least favorable (precise) model of* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *for the set of loss functions*

$$\mathcal{W} \;=\; \big\{W^{(\pi)} \;\big|\; \pi \in \mathcal{P}\big\}$$

where $W^{(\pi)}$ *denotes the loss function*

$$W^{(\pi)}: \;\; (\theta, t) \;\mapsto\; n\pi_\theta W_\theta(t)$$

The next theorem shows how least favorable models can be used to deal with situations where the distribution of the data as well as the prior is assumed to be imprecise. A decision procedure, i.e. a generalized randomization, is optimal if it minimizes the upper Bayes risk

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \sigma, (W_\theta)_\theta\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{Q}_\theta)_\theta, \sigma, (W_\theta)_\theta\big)$$

In case of a precise model $(Q_\theta)_{\theta\in\Theta}$, the upper Bayes risk is

$$R_{\overline{\Pi}}\big((Q_\theta)_\theta, \sigma, W\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((Q_\theta)_\theta, \sigma, W\big)$$

Theorem 4.15 *If* $(Q_\theta)_{\theta\in\Theta}$ *is a simultaneously least favorable model of* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *for* $\mathcal{P}$, *there is a generalized randomization* $\tilde\sigma \in \mathcal{T}(\mathcal{Y}, \mathbb{D})$ *which minimizes*

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \qquad \textit{and also} \qquad R_{\overline{\Pi}}\big((Q_\theta)_\theta, \sigma, W\big)$$

over $\mathcal{T}(\mathcal{Y}, \mathbb{D})$.

The proof is essentially an application of the following lemma which is based on the minimax theorem. Again, compactness of the credal set is crucial.

Lemma 4.16

$$\textbf{(a)}\quad \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R_\pi\big(((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

$$\textbf{(b)}\quad \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_\pi\big(((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

$$\textbf{(c)}\quad \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big(((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

Proof: The three statements (a), (b) and (c) are proven simultaneously. Therefore, let

$$\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \;\in\; \big\{\mathcal{T}_r(\mathcal{Y},\mathbb{D})\,,\ \mathcal{T}_0(\mathcal{Y},\mathbb{D})\,,\ \mathcal{T}(\mathcal{Y},\mathbb{D})\big\}$$

be fixed in the following.

Put

$$\Gamma(\sigma,\pi) \;=\; R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \qquad \forall\,\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})\,,\ \ \forall\,\pi\in\mathcal{P}$$

With respect to the $\mathcal{L}_\infty(\Theta,2^\Theta)$ - topology on $\mathrm{ba}(\Theta,2^\Theta)$, the credal set $\mathcal{P}$ is compact (Corollary 2.16) and the map

$$\pi \;\mapsto\; \Gamma(\sigma,\pi) \;=\; \int_\Theta \sup_{Q_\theta\in\mathcal{M}_\theta} \sigma(Q_\theta)[W_\theta]\,\pi(d\theta)$$

is continuous and concave for every $\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})$. Furthermore,

$$\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \;\to\; \mathbb{R}\,,\qquad \sigma\;\mapsto\;\Gamma(\sigma,\pi)$$

is convex for every $\pi\in\mathcal{P}$. Then, it follows from (Fan, 1953, Theorem 2) that

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})}\ \sup_{\pi\in\mathcal{P}} \Gamma(\sigma,\pi) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} \Gamma(\sigma,\pi) \tag{4.17}$$

Finally, (a), (b) and (c) follow from (4.17) because

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\Gamma(\sigma,\pi)$$

where $\mathcal{P}$ denotes the credal set of $\overline{\Pi}$. □

Proof of Theorem 4.15: Note that the precise model $(Q_\theta)_{\theta\in\Theta}$ is a special case of an imprecise model. Hence, Lemma 4.16 is also applicable for $(Q_\theta)_{\theta\in\Theta}$ instead of $(\overline{Q}_\theta)_{\theta\in\Theta}$.

Consequently, a twofold application of Lemma 4.16 and simultaneous least favorability yield

$$\begin{aligned}\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) &= \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\\ &= \sup_{\pi\in\mathcal{P}}\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((Q_\theta)_\theta,\sigma,W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((Q_\theta)_\theta,\sigma,W\big)\end{aligned} \tag{4.18}$$

Lower semicontinuity of

$$\sigma \;\mapsto\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

and compactness of $\mathcal{T}(\mathcal{Y},\mathbb{D})$ ensure existence of a minimum $\tilde{\sigma}$ (cf. (Denkowski et al., 2003, Theorem 1.3.11)). Additionally,

$$R_{\overline{\Pi}}\big((Q_\theta)_\theta,\tilde{\sigma},W\big) \;\le\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\tilde{\sigma},W\big) = \\ = \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \overset{(4.18)}{=} \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((Q_\theta)_\theta,\sigma,W\big)$$

□

Remark 4.17 *It can easily be read off from the above proof that a generalized randomization $\tilde{\sigma}$ which minimizes $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$ minimizes $R_{\overline{\Pi}}\big((Q_\theta)_\theta,\sigma,W\big)$, too. However, the reverse statement will not always be true.*[3] *So, it does not suffice to find a generalized randomization $\hat{\sigma}$ which minimizes $R_{\overline{\Pi}}\big((Q_\theta)_\theta,\sigma,W\big)$. It still has to be checked that $\hat{\sigma}$ really minimizes $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$. Theorem 4.15 only states that there is a generalized randomization which solves both minimization problems.*

Theorem 4.15 is stated in terms of *generalized* randomization. The following proposition enables us to formulate this result also in terms of ordinary randomizations.

Proposition 4.18 *There is a generalized randomization $\tilde{\sigma}\in\mathcal{T}(\mathcal{Y},\mathbb{D})$ such that*

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\tilde{\sigma},W\big) \;=\; \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big)$$

Furthermore,

$$\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,W\big)$$

coincides for $\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \,=\, \mathcal{T}_r(\mathcal{Y},\mathbb{D})\,,\;=\,\mathcal{T}_0(\mathcal{Y},\mathbb{D})$ and $=\,\mathcal{T}(\mathcal{Y},\mathbb{D})\,$.

Proof: Note that

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

where $\mathcal{P}$ denotes the credal set of $\overline{\Pi}$. According to the proof of Proposition 4.5, the map

$$\mathcal{T}(\mathcal{Y},\mathbb{D}) \;\to\; \mathbb{R}\,, \qquad \sigma \;\mapsto\; R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

is lower semicontinuous for every $\pi\in\mathcal{P}$. This implies that

$$\sigma \;\mapsto\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

is the supremum of lower semicontinuous functions and, therefore, is also lower semicontinuous. Next, compactness of $\mathcal{T}(\mathcal{Y},\mathbb{D})$ ensure existence of some $\tilde{\sigma}$ which minimizes $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$ (cf. (Denkowski et al., 2003, Theorem 1.3.11)).

For the proof of the second statement, let

$$\mathcal{T}_*(\mathcal{Y},\mathbb{D}) \;\in\; \big\{\mathcal{T}_r(\mathcal{Y},\mathbb{D})\,,\;\mathcal{T}_0(\mathcal{Y},\mathbb{D})\,,\;\mathcal{T}(\mathcal{Y},\mathbb{D})\big\}$$

be fixed in the following. Then, Proposition 4.5 and a twofold application of Lemma 4.16 imply

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\;\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;= \\ =\; \sup_{\pi\in\mathcal{P}}\;\inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,W\big) \;=\; \inf_{\sigma\in\mathcal{T}_*(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,W\big)$$

□

[3] In case of hypothesis testing, for example, this follows from (Augustin, 1998, p. 162ff).

Proposition 4.18 is the generalization of Proposition 4.5 for imprecise prior distributions. As a consequence, it does not matter if we consider ordinary or generalized randomizations in decision theory where the prior distribution and the distribution of the data may be imprecise.

So, we can formulate the above theorem also in terms of ordinary randomizations.

Theorem 4.19 *Let* $(Q_\theta)_{\theta\in\Theta}$ *be a simultaneously least favorable model of* $(\mathcal{M}_\theta)_{\theta\in\Theta}$ *for* $\mathcal{P}$ *and* $\varepsilon > 0$*. Then, there is an ordinary randomization* $\tilde{\sigma}_0 \in \mathcal{T}_0(\mathcal{Y},\mathbb{D})$ *such that*

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \tilde{\sigma}_0, W\big) \;\le\; \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon \tag{4.19}$$

and also

$$R_{\overline{\Pi}}\big((Q_\theta)_\theta, \tilde{\sigma}_0, W\big) \;\le\; \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((Q_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon \tag{4.20}$$

Proof: Proposition 4.18 ensures existence of an ordinary randomization $\tilde{\sigma}_0 \in \mathcal{T}_0(\mathcal{Y},\mathbb{D})$ which fulfills (4.19). Analogously to the proof of Theorem 4.15,

$$\begin{aligned}
R_{\overline{\Pi}}\big((Q_\theta)_\theta, \tilde{\sigma}_0, W\big) \;\le\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \tilde{\sigma}_0, W\big) &\overset{(4.19)}{\le} \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon \\
\overset{\text{Prop. 4.18}}{=} \quad & \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon \;= \\
\overset{(4.18)}{=} \quad & \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((Q_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon \;= \\
\overset{\text{Prop. 4.18}}{=} \quad & \inf_{\sigma\in\mathcal{T}_0(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((Q_\theta)_\theta, \sigma, W\big) \,+\, \varepsilon
\end{aligned}$$

□

4.3 Statistical hypothesis testing

4.3.1 Decision theoretic formulation of hypothesis testing

Let us first consider simple hypothesis testing

$$P_0 \qquad \text{vs.} \qquad P_1$$

where P_0 and P_1 are probability measures on some measurable space $(\mathcal{X},\mathcal{A})$.

An $\mathcal{A}$ - *measurable test* is a map $\varphi : \mathcal{X} \longrightarrow [0,1]$ which is measurable with respect to the σ-algebra $\mathcal{A}$ on $\mathcal{X}$ and the Borel-σ-algebra on $[0,1]$, i.e.:

$$\varphi \,\in\, \mathcal{L}_\infty(\mathcal{X},\mathcal{A}), \qquad 0 \,\le\, \varphi \,\le\, 1$$

Let

$$\mathcal{T}_0 \;=\; \big\{\varphi \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \;\big|\; 0 \le \varphi \le 1\big\}$$

denote the set of all these tests. Of course, this definition depends on the fixed $(\mathcal{X},\mathcal{A})$ although it is not made visible in this notation.

A test $\tilde{\varphi}$ is called *optimal* if it solves the minimization problem

$$P_1[1-\varphi] \;=\; \min_{\varphi}! \qquad \varphi \in \mathcal{T}_0\,, \quad P_0[\varphi] \;\leq\; \alpha \tag{4.21}$$

where $\alpha \in [0,1]$ is some fixed significance level.

It is well known how this simple hypothesis testing fits into the decision theoretic framework:
Put $\Theta = \{0,1\}$ as the set of all states of nature. Then, (P_0, P_1) is a precise model on $(\mathcal{X}, \mathcal{A})$. Hypothesis testing means to decide if the null hypothesis is rejected or not. So, the possible decisions are

$$\begin{array}{ll} d \;=\; 0: & \text{do not reject the null hypothesis} \\ d \;=\; 1: & \text{reject the null hypothesis} \end{array}$$

That is, the decision space is $(\mathbb{D}, \mathcal{D}) \;=\; \big(\{0,1\}, 2^{\{0,1\}}\big)$.
The loss function W is defined by

$$W_0 \;=\; I_{\{1\}}\,, \qquad W_1 \;=\; I_{\{0\}}$$

where $I_{\{k\}} : \mathbb{D} \to \mathbb{R}$ is the indicator function. That is, a wrong decision leads to the loss 1 and a correct decision leads to the loss 0.

Every test $\varphi \in \mathcal{T}_0$ defines a Markov kernel

$$\mathcal{X} \times \mathcal{D} \;\to\; \mathbb{R}\,, \qquad (x, D) \;\mapsto\; \mathrm{Bin}\big(1, \varphi(x)\big)(D)$$

Conversely, every Markov kernel $\tau : \mathcal{X} \times \mathcal{D} \to \mathbb{R}\,,\ (x,D) \mapsto \tau_x(D)$ defines a test $\varphi \in \mathcal{T}_0$ via

$$\varphi(x) \;=\; \tau_x\big(\{1\}\big) \tag{4.22}$$

In this way, the set of all tests $\mathcal{T}_0$ corresponds to the set of all ordinary randomizations $\mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$. This justifies the similarity of the notation. Let φ be a test and σ its corresponding ordinary randomization. Then, it is easy to see that

$$\sigma(\mu)[I_{\{1\}}] \;=\; \int_{\mathcal{X}}\int_{\mathbb{D}} I_{\{1\}}(t)\tau_x(dt)\mu(dx) \;\overset{(4.22)}{=}\; \mu[\varphi]$$

and

$$\sigma(\mu)[I_{\{0\}}] \;=\; \int_{\mathcal{X}}\int_{\mathbb{D}} I_{\{0\}}(t)\tau_x(dt)\mu(dx) \;\overset{(4.22)}{=}\; \mu[1-\varphi]$$

for every $\mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A})$.

According to Section 3.2 and Subsection 3.3.1, the risk function of an ordinary randomization σ defined by a Markov kernel τ is

$$\{0,1\} \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; \sigma(P_\theta)[W_\theta] \;=\; \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\tau_x(dt)P_\theta(dx)$$

Note that

$$\sigma(P_0)[W_0] \;=\; \int_{\mathcal{X}}\int_{\mathbb{D}} W_0(t)\tau_x(dt)P_0(dx) \;=\; P_0[\varphi]$$

is the type I error and

$$\sigma(P_1)[W_1] \;=\; \int_{\mathcal{X}}\int_{\mathbb{D}} W_1(t)\tau_x(dt)P_1(dx) \;=\; 1 - P_1[\varphi] \;=\; P_1[1-\varphi]$$

is the type II error of the test φ which corresponds to σ and τ.

Let π be any prior distribution on $\Theta = \{0,1\}$; put

$$\pi_0 \;=\; \pi[I_{\{0\}}] \qquad \text{and} \qquad \pi_1 \;=\; \pi[I_{\{1\}}]$$

According to Section 3.2, an ordinary randomization $\tilde{\sigma} \in \mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$ is optimal if it minimizes the Bayes risk

$$R_\pi\big((P_0,P_1),\sigma,W\big) \;=\; \pi_0\sigma(P_0)[W_0] + \pi_1(P_1)[W_1]$$

Let $\sigma \in \mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$ be an ordinary randomization and $\varphi \in \mathcal{T}_0$ its corresponding test, then the Bayes risk of σ is equal to

$$R_\pi\big((P_0,P_1),\sigma,W\big) \;=\; \pi_0 P_0[\varphi] + \pi_1 P_1[1-\varphi]$$

It is well known that the prior π can be chosen so that the corresponding decision problem is equivalent to the original testing problem. That is: Testing problem (4.21) can be solved by solving an appropriate decision problem.

The same is true for hypothesis testing where each hypothesis consists of an imprecise probability, i.e. a coherent upper prevision; cf. Theorem 4.20. The present subsection explaines how this "imprecise testing problem" can be formulated by decision theory. But first of all, we have to take a closer look at the testing problem.

Now, $\mathcal{X}$ is a set and $\mathcal{A}$ is an algebra on $\mathcal{X}$ and we consider imprecise simple hypothesis testing

$$\overline{P}_0 \qquad \text{vs.} \qquad \overline{P}_1 \tag{4.23}$$

where $\overline{P}_0$ and $\overline{P}_1$ are coherent upper previsions on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$.
Again,

$$\mathcal{T}_0 \;=\; \big\{\varphi \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \;\big|\; 0 \le \varphi \le 1\big\}$$

denotes the set of all $\mathcal{A}$-measurable tests.

A test $\tilde{\varphi}$ is called *optimal* if it solves the minimization problem

$$\overline{P}_1[1-\varphi] \;=\; \min_{\varphi}! \qquad \varphi \in \mathcal{T}_0\,, \quad \overline{P}_0[\varphi] \;\le\; \alpha \tag{4.24}$$

where $\alpha \in [0,1]$ is some fixed significance level.

This formulation of the optimization problem is adequate and coincides with the optimization problem of a testing problem in classical statistics:

Let $\mathcal{M}_0$ be the credal set of $\overline{P}_0$ and let $\mathcal{M}_1$ be the credal set of $\overline{P}_1$ on $(\mathcal{X},\mathcal{A})$. Consider testing problem

$$\mathcal{M}_0 \qquad \text{vs.} \qquad \mathcal{M}_1 \tag{4.25}$$

This is a common testing problem in classical statistics[4] where only precise probabilities are involved. Within this classical setup, a test $\tilde{\varphi}$ is usually called *optimal* with respect to the minimax criterion if it solves minimization problem

$$\sup_{P_1 \in \mathcal{M}_1} P_1[1-\varphi] \;=\; \min_{\varphi}! \qquad \varphi \in \mathcal{T}_0\,, \quad \sup_{P_0 \in \mathcal{M}_0} P_0[\varphi] \;\leq\; \alpha \tag{4.26}$$

That is, an optimal test minimizes the maximal type II error. As in Chapter 3, this is again a worst case consideration. Obviously, minimization problem (4.24) is equal to minimization problem (4.26). That is, imprecise simple hypothesis testing (4.23) coincides with the classical testing problem (4.25) and the solution of (4.23) is an ordinary minimax test.

The standard reference for the classical testing problem (4.25) is Baumann (1968). The importance of Baumann's results for testing between imprecise probabilities in case of F-probabilities was discovered by Augustin (1998); cf. also Augustin (2002). Furthermore, (Augustin, 1998, § 3.1 and p. 121–123) contains a detailed discussion of the connections between F-probabilities and the classical testing problem (4.25).

In the following, it is explained in detail how imprecise simple hypothesis testing (4.23) fits into the decision theoretic framework presented in this book:

Again, $\Theta = \{0,1\}$ is the set of all states of nature, $(\overline{P}_0, \overline{P}_1)$ is an imprecise model on $(\mathcal{X}, \mathcal{A})$ and the possible decisions are

$$\begin{array}{lll} d \;=\; 0 \;: & & \text{do not reject the null hypothesis} \\ d \;=\; 1 \;: & & \text{reject the null hypothesis} \end{array}$$

That is, the decision space is $(\mathbb{D}, \mathcal{D}) \;=\; \big(\{0,1\}, 2^{\{0,1\}}\big)$. The loss function W is again defined by

$$W_0 \;=\; I_{\{1\}}\,, \qquad W_1 \;=\; I_{\{0\}}$$

where $I_{\{k\}} : \mathbb{D} \to \mathbb{R}$ is the indicator function.

As stated before, the set of all tests $\mathcal{T}_0$ corresponds to the set of all ordinary randomizations $\mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$ and, according to Section 3.2 and Subsection 3.3.1, the risk function of an ordinary randomization σ defined by a Markov kernel τ is

$$\{0,1\} \;\to\; \mathbb{R}\,, \qquad \theta \;\mapsto\; \sup_{P_\theta \in \mathcal{M}_\theta} \sigma(P_\theta)[W_\theta] \;=\; \sup_{P_\theta \in \mathcal{M}_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\tau_x(dt)P_\theta(dx)$$

Here,

$$\sup_{P_0 \in \mathcal{M}_0} \sigma(P_0)[W_0] \;=\; \sup_{P_0 \in \mathcal{M}_0} \int_{\mathcal{X}} \int_{\mathbb{D}} W_0(t)\tau_x(dt)P_0(dx) \;=\; \overline{P}_0[\varphi]$$

is the type I error and

$$\sup_{P_1 \in \mathcal{M}_1} \sigma(P_1)[W_1] \;=\; \sup_{P_1 \in \mathcal{M}_1} \int_{\mathcal{X}} \int_{\mathbb{D}} W_1(t)\tau_x(dt)P_1(dx) \;=\; \overline{P}_1[1-\varphi]$$

is the type II error of the test φ which corresponds to σ and τ .

[4] appart from σ-additivity, of course

With the above settings, a randomizations $\sigma \in \mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$ is called *level-α-randomization* if its corresponding test $\varphi \in \mathcal{T}_0$ is a level-α-test – that is, if

$$\sup_{P_0 \in \mathcal{M}_0} \sigma(P_0)[W_0] \;\le\; \alpha$$

As in case of classical simple hypothesis testing we have the following result:

Theorem 4.20 *There is a prior π such that the following is true:*
A test $\tilde{\varphi}$ with type I error [5]

$$\overline{P}_0[\tilde{\varphi}] \;=\; \alpha \tag{4.27}$$

solves testing problem (4.24) if and only if its corresponding randomization $\tilde{\sigma}$ solves the corresponding decision problem – that is, if it minimizes the Bayes risk

$$R_\pi\big((\overline{P}_0, \overline{P}_1), \sigma, W\big) \;=\; \sum_{\theta \in \{0,1\}} \pi_\theta \cdot \sup_{P_\theta \in \mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]\,, \qquad \sigma \;\in\; \mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$$

The proof of this statement is postponed to the following subsection where an analogous result is proven in case of "generalized tests". These generalized tests, which are investigated in the following subsection, are the corresponding counterpart of generalized randomizations. Theorem 4.20 will follow from the results of Subsection 4.3.2 as an easy corollary.

4.3.2 Generalized tests

It is well known that there does not need to be an optimal test which solves testing problem (4.26) i.e. (4.24). However, testing problems can be rewritten into decision problems, and we already know from Proposition 4.5 that every decision problem can be solved by a *generalized* randomization. This is one reason why generalized tests are introduced in this subsection. Generalized tests are the corresponding counterparts of generalized randomization. It is shown that there is always a generalized level-α-test which is optimal in testing problem (4.23).

Let $\mathcal{X}$ be again a set with algebra $\mathcal{A}$ and

$$\mathcal{T}_0 \;=\; \big\{\varphi \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}) \;\big|\; 0 \le \varphi \le 1\big\}$$

denote the set of all $\mathcal{A}$-measurable tests. As described in the previous subsection, the ordinary randomizations

$$\sigma \;:\; \mathrm{ba}(\mathcal{X}, \mathcal{A}) \;\to\; \mathrm{ba}\big(\{0,1\}, 2^{\{0,1\}}\big)\,, \qquad \sigma \in \mathcal{T}_0\big(\mathcal{X}, \{0,1\}\big)$$

correspond to tests $\varphi \in \mathcal{T}_0$ via

$$\sigma(\mu)[I_{\{1\}}] \;=\; \mu[\varphi]\,, \qquad \sigma(\mu)[I_{\{0\}}] \;=\; \mu[1-\varphi] \qquad \forall\, \mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A})$$

In the same way, generalized randomizations correspond to "generalized tests" which can be defined in the following way:

[5] Equation (4.27) is not a strong assumption because every optimal level-α-test can be transformed into an optimal test which fulfills (4.27).

Definition 4.21 *A* generalized test on $(\Omega, \mathcal{A})$ *is a map*

$$\phi : \ \mathrm{ba}(\Omega, \mathcal{A}) \ \rightarrow \ \mathbb{R}$$

which is linear and $\phi(P) \ \in \ [0,1] \ \ \forall\, P \in \mathrm{ba}_1^+(\Omega, \mathcal{A})$ *.*
The elements of $\mathcal{T}_0$ *are called* ordinary test on $(\Omega, \mathcal{A})$ *. The set of all generalized tests on* $(\Omega, \mathcal{A})$ *is denoted by* $\mathcal{T}$*.*

Note that every ordinary test φ defines a generalized test ϕ via

$$\phi(\mu) \ = \ \mu[\varphi] \qquad \forall\, \mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A})$$

Therefore, we may also write

$$\varphi(\mu) := \mu[\varphi] \qquad \text{and} \qquad \mathcal{T}_0 \ \subset \ \mathcal{T} \tag{4.28}$$

Generalized randomizations

$$\sigma \ : \ \mathrm{ba}(\mathcal{X}, \mathcal{A}) \ \rightarrow \ \mathrm{ba}\big(\{0,1\}, 2^{\{0,1\}}\big)\,, \qquad \sigma \in \mathcal{T}\big(\mathcal{X}, \{0,1\}\big)$$

correspond to tests $\phi \in \mathcal{T}$ via

$$\sigma(\mu)[I_{\{1\}}] \ = \ \phi(\mu)\,, \qquad \sigma(\mu)[I_{\{0\}}] \ = \ \mu[I_{\mathcal{X}}] - \phi(\mu) \qquad \forall\, \mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A})$$

The dependencies are illustrated by the following picture:

$$\begin{tikzcd}
\text{ordinary tests} \arrow[r, "\text{correspond to}"] \arrow[d, "\text{generalize}"] & \text{ordinary randomizations} \arrow[d, "\text{generalize}"] \\
\text{generalized tests} \arrow[r, "\text{correspond to}"] & \text{generalized randomizations}
\end{tikzcd}$$

As in the previous subsection, we consider the testing problem

$$\overline{P}_0 \qquad \text{vs.} \qquad \overline{P}_1$$

where each $\overline{P}_i$ is a coherent upper prevision with credal set $\mathcal{M}_i$. The (supremal) type I and type II errors of a generalized test $\phi \in \mathcal{T}$ may be defined by

$$\text{type I error:} \quad \sup_{P_0 \in \mathcal{M}_0} \ \phi(P_0)\ , \qquad \text{type II error:} \quad \sup_{P_1 \in \mathcal{M}_1} \ 1 - \phi(P_1) \tag{4.29}$$

In case of ordinary randomizations these error terms simplify to the ordinary ones if ϕ is given by an ordinary test. The minimization problem is

$$\sup_{P_1 \in \mathcal{M}_1} \ 1 - \phi(P_1) \ = \ \min_{\phi}! \qquad \phi \in \mathcal{T}\,, \quad \sup_{P_0 \in \mathcal{M}_0} \ \phi(P_0) \ \leq \ \alpha \tag{4.30}$$

then.
In addition, the above settlings fit into the decision theoretic formalization of hypothesis testing described in the previous subsection. In order to see this, let again $\Theta = \{0,1\}$ be the set of all states of nature and $(\mathbb{D}, \mathcal{D}) \ = \ \big(\{0,1\}, 2^{\{0,1\}}\big)$ the decision space. $(\overline{P}_0, \overline{P}_1)$ is an imprecise model on $(\mathcal{X}, \mathcal{A})$; the loss function W is defined by

$$W_0 \ = \ I_{\{1\}}\,, \qquad W_1 \ = \ I_{\{0\}}$$

Let $\phi \in \mathcal{T}$ be a generalized test and $\sigma \in \mathcal{T}\big(\mathcal{X},\{0,1\}\big)$ its corresponding generalized randomization. Then,

$$\sup_{P_0\in\mathcal{M}_0} \sigma(P_0)[W_0] \;=\; \sup_{P_0\in\mathcal{M}_0} \phi(P_0) \tag{4.31}$$

is the type I error and

$$\sup_{P_1\in\mathcal{M}_1} \sigma(P_1)[W_1] \;=\; \sup_{P_1\in\mathcal{M}_1} 1-\phi(P_1) \tag{4.32}$$

is the type II error of the generalized test ϕ – just as in case of ordinary tests in Subsection 4.3.1.

The main goal of the rest of the present subsection is to show that it does not matter if ordinary or generalized tests are used in testing problem (4.23).
To this end, endow $\mathcal{T}$ with the topology of pointwise convergence on $\mathrm{ba}(\Omega,\mathcal{A})$. It is the smallest topology so that

$$\mathcal{T} \;\to\; \mathbb{R}, \qquad \phi \;\mapsto\; \phi(\mu)$$

is continuous for every $\mu \in \mathrm{ba}(\Omega,\mathcal{A})$. It is easy to see that this topology corresponds to the topology of pointwise convergence on the set of all generalized randomizations $\mathcal{T}\big(\mathcal{X},\{0,1\}\big)$:
Let $(\phi_\beta)_{\beta\in B}$ be a net of generalized tests and $\phi \in \mathcal{T}$ be another generalized test; let $(\sigma_\beta)_{\beta\in B}$ and σ be the corresponding generalized randomizations. Then,

$$\phi_\beta \;\underset{\beta}{\longrightarrow}\; \phi \qquad \text{if and only if} \qquad \sigma_\beta \;\underset{\beta}{\longrightarrow}\; \sigma$$

This implies the following theorem which corresponds to Theorem 3.9 and Theorem 3.10:

Theorem 4.22 *$\mathcal{T}$ is compact and $\mathcal{T}_0$ is dense in $\mathcal{T}$.*

The set of the ordinary level-α-tests in testing problem (4.23) is denoted by

$$\mathcal{T}_0(\alpha) \;=\; \Big\{\varphi \in \mathcal{T}_0 \,\Big|\, \sup_{P_0\in\mathcal{M}_0} P_0[\varphi] \;\le\; \alpha\Big\}$$

and the set of the generalized level-α-tests is denoted by

$$\mathcal{T}(\alpha) \;=\; \Big\{\phi \in \mathcal{T} \,\Big|\, \sup_{P_0\in\mathcal{M}_0} \phi(P_0) \;\le\; \alpha\Big\}$$

The following theorem is the analog of Theorem 4.20 in case of generalized tests:

Theorem 4.23 *There is a prior π such that the following is true:*
A generalized test $\tilde{\phi}$ with type I error

$$\sup_{P_0\in\mathcal{M}_0} \tilde{\phi}(P_0) \;=\; \alpha \tag{4.33}$$

solves testing problem (4.30) if and only if its corresponding randomization $\tilde{\sigma}$ solves the corresponding decision problem – that is, if it minimizes the Bayes risk

$$R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big) \;=\; \sum_{\theta\in\{0,1\}} \pi_\theta \cdot \sup_{P_\theta\in\mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]\;, \qquad \sigma \;\in\; \mathcal{T}\big(\mathcal{X},\{0,1\}\big)$$

Proof: It follows from Lemma 4.24 below that there is some $a \in [0, \infty)$ such that

$$\begin{aligned} \inf_{\phi \in \mathcal{T}(\alpha)} \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;+\; a \cdot \alpha \;\; &= \\ = \inf_{\phi \in \mathcal{T}} \Big(\sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;+\; a \cdot \sup_{P_0 \in \mathcal{M}_0} \phi(P_0) \Big) & \qquad (4.34) \end{aligned}$$

Let ϕ be a generalized test and σ its corresponding generalized randomization. Then, the above settings imply

$$\begin{aligned} \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;+\; a \cdot \sup_{P_0 \in \mathcal{M}_0} \phi(P_0) \;\; &= \\ = \sup_{P_1 \in \mathcal{M}_1} \sigma(P_1)[W_1]\big) \;+\; a \cdot \sup_{P_0 \in \mathcal{M}_0} \sigma(P_0)[W_0] \;\; &= \\ = \;\; (1+a) \cdot \sum_{\theta \in \{0,1\}} \pi_\theta \cdot \sup_{P_\theta \in \mathcal{M}_\theta} \sigma(P_\theta)[W_\theta] \;\; &= \\ = \;\; (1+a) \cdot R_\pi\big((\overline{P}_0, \overline{P}_1), \sigma, W\big) & \qquad (4.35) \end{aligned}$$

where

$$\pi_0 \;=\; \frac{a}{1+a} \qquad \text{and} \qquad \pi_1 \;=\; \frac{1}{1+a}$$

Let $\tilde{\phi}$ be a level-α-test which fulfills (4.33) and let $\tilde{\sigma}$ be its corresponding randomization. Firstly, asume that $\tilde{\phi}$ solves testing problem (4.24). Then,

$$\begin{aligned} R_\pi\big((\overline{P}_0, \overline{P}_1), \tilde{\sigma}, W\big) \;\overset{(4.35)}{=}\; \frac{1}{1+a} \Big(\sup_{P_1 \in \mathcal{M}_1} \tilde{\phi}(P_1) \;+\; a \cdot \sup_{P_0 \in \mathcal{M}_0} \tilde{\phi}(P_0) \Big) \;\; &= \\ \overset{(4.33)}{=} \; \frac{1}{1+a} \Big(\sup_{P_1 \in \mathcal{M}_1} \tilde{\phi}(P_1) \;+\; a \cdot \alpha \Big) \;\; &= \\ = \;\; \frac{1}{1+a} \Big(\inf_{\phi \in \mathcal{T}(\alpha)} \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;+\; a \cdot \alpha \Big) \;\; &= \\ \overset{(4.34),(4.35)}{=} \; \inf_{\sigma \in \mathcal{T}} R_\pi\big((\overline{P}_0, \overline{P}_1), \sigma, W\big) & \end{aligned}$$

Conversely, asume that $\tilde{\sigma}$ minimizes the Bayes risk in (4.35). Then,

$$\begin{aligned} \inf_{\phi \in \mathcal{T}(\alpha)} \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;=\; \inf_{\phi \in \mathcal{T}(\alpha)} \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) \;+\; a \cdot \alpha \;-\; a \cdot \alpha \;\; &= \\ \overset{(4.34),(4.35)}{=} \; (1+a) \cdot \inf_{\sigma \in \mathcal{T}} R_\pi\big((\overline{P}_0, \overline{P}_1), \sigma, W\big) \;-\; a \cdot \alpha \;\; &= \\ = \; (1+a) \cdot R_\pi\big((\overline{P}_0, \overline{P}_1), \tilde{\sigma}, W\big) \;-\; a \cdot \alpha \;\; \overset{(4.35),(4.33)}{=} \;\; \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1) & \end{aligned}$$

□

The following lemma can be formulated for ordinary and generalized tests – recall the notation introduced in (4.28).

Lemma 4.24 *In case of*

$$\mathcal{T}_* \;=\; \mathcal{T}_0\,, \qquad \mathcal{T}_*(\alpha) \;=\; \mathcal{T}_0(\alpha)$$

and also in case of

$$\mathcal{T}_{*} \;=\; \mathcal{T}\,, \qquad \mathcal{T}_{*}(\alpha) \;=\; \mathcal{T}(\alpha)$$

there is some $a \in [0,\infty)$ *such that*

$$\begin{aligned}\inf_{\phi\in\mathcal{T}_*(\alpha)}\;\sup_{P_1\in\mathcal{M}_1}\big(1-\phi(P_1)\big)\;+\;a\cdot\alpha \;&=\\ =\;\inf_{\phi\in\mathcal{T}_*}\bigg(\sup_{P_1\in\mathcal{M}_1}\big(1-\phi(P_1)\big)\;+\;a\cdot\sup_{P_0\in\mathcal{M}_0}\phi(P_0)\bigg)& \qquad (4.36)\end{aligned}$$

Proof: In testing problem (4.23), we consider minimization problem

$$\sup_{P_1\in\mathcal{M}_1} 1-\phi(P_1) \;=\; \inf_{\phi}! \qquad \phi\in\mathcal{T}_*\,, \quad \sup_{P_0\in\mathcal{M}_0}\phi(P_0)\;\le\;\alpha$$

where $\phi \;\mapsto\; \sup_{P_1\in\mathcal{M}_1} 1-\phi(P_1)$ and $\phi \;\mapsto\; \sup_{P_0\in\mathcal{M}_0}\phi(P_0)$ are convex functions on $\mathcal{T}_*$, and $\mathcal{T}_*$ is a convex subset of a real topological vector space.

Then, it follows from Rieder (1994, Theorem B.2.1) that there is some $a\in[0,\infty)$ such that

$$\begin{aligned}\inf_{\phi\in\mathcal{T}_*(\alpha)}\;\sup_{P_1\in\mathcal{M}_1}\big(1-\phi(P_1)\big)\;+\;a\cdot\alpha \;&=\\ =\;\inf_{\phi\in\mathcal{T}_*}\bigg(\sup_{P_1\in\mathcal{M}_1}\big(1-\phi(P_1)\big)\;+\;a\cdot\sup_{P_0\in\mathcal{M}_0}\phi(P_0)\bigg)& \qquad (4.37)\end{aligned}$$

□

The following theorem says that it does not matter whether only ordinary tests are permitted or generalised tests are also permitted.

Theorem 4.25 *The infimal type II error over all generalized level-α-tests coincides with the infimal type II error over all ordinary level-α-tests in testing problem (4.23):*

$$\inf_{\phi\in\mathcal{T}(\alpha)}\;\sup_{P_1\in\mathcal{M}_1} 1-\phi(P_1) \quad=\quad \inf_{\varphi\in\mathcal{T}_0(\alpha)}\;\sup_{P_1\in\mathcal{M}_1} 1-\varphi(P_1) \qquad (4.38)$$

Furthermore, there is always an optimal *generalized level-α-test which achieves the infimum in (4.38).*

Equation (4.38) does not easily follow from the fact that $\mathcal{T}_0$ is dense in $\mathcal{T}$. This is because it is not clear if $\mathcal{T}_0(\alpha)$ is dense in $\mathcal{T}(\alpha)$. This problem is avoided in the following proof by using Lagrange techniques (Lemma 4.24). That way we turn over to a corresponding decision problem. Due to Proposition 4.5, we already know that it makes no difference in decision problems whether generalized randomizations are permitted or not. [6]

Proof: Since $\mathcal{T}_0(\alpha)\;\subset\;\mathcal{T}(\alpha)$, it is enough to show

$$\sup_{P_1\in\mathcal{M}_1} 1-\phi(P_1) \quad\ge\quad \inf_{\varphi\in\mathcal{T}_0(\alpha)}\;\sup_{P_1\in\mathcal{M}_1} 1-\varphi(P_1) \qquad \forall\,\phi\in\mathcal{T}(\alpha)$$

[6] Though the proof of Theorem 4.25 is very similar to the proof of Theorem 4.20, the statement of the latter theorem cannot be used here. It would have been possible to formulate Theorem 4.20 in such a way that Theorem 4.25 was a simple corollary of Theorem 4.20. However, this has not been done because this would have obscured the real intent of Theorem 4.20.

According to Lemma 4.24, there is a Lagrange multiplier $a \in [0,\infty)$ so that

$$\begin{aligned}\inf_{\varphi\in\mathcal{T}_0(\alpha)}\ \sup_{P_1\in\mathcal{M}_1}\big(1-\varphi(P_1)\big)\ +\ a\cdot\alpha\ &=\\ =\ \inf_{\varphi\in\mathcal{T}_0}\left(\sup_{P_1\in\mathcal{M}_1}\big(1-\varphi(P_1)\big)\ +\ a\cdot\sup_{P_0\in\mathcal{M}_0}\varphi(P_0)\right)\end{aligned} \tag{4.39}$$

Fix any $\phi\in\mathcal{T}(\alpha)$. A suitable convex combination of ϕ and

$$\phi_1:\quad \mathrm{ba}(\Omega,\mathcal{A})\ \to\ \mathbb{R}\,,\qquad \mu\ \mapsto\ \mu[I_\Omega]$$

yields some $\hat\phi\in\mathcal{T}(\alpha)$ so that

$$\sup_{P_1\in\mathcal{M}_1}1-\hat\phi(P_1)\ \le\ \sup_{P_1\in\mathcal{M}_1}1-\phi(P_1)\,,\qquad \sup_{P_0\in\mathcal{M}_0}\hat\phi(P_0)\ =\ \alpha \tag{4.40}$$

Put the index set Θ, the prior π and the loss function W as in the proof of Theorem 4.23. As in (4.35), it follows that

$$\sup_{P_1\in\mathcal{M}_1}\big(1-\phi(P_1)\big)+a\cdot\sup_{P_0\in\mathcal{M}_0}\phi(P_0)\ =\ (1+a)\cdot R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big) \tag{4.41}$$

for every generalized test ϕ with corresponding generalized randomization σ. Finally,

$$\begin{aligned}\sup_{P_1\in\mathcal{M}_1}1-\phi(P_1)\ &\overset{(4.40)}{\ge}\ \sup_{P_1\in\mathcal{M}_1}1-\hat\phi(P_1)\ =\\ &\overset{(4.40)}{=}\ \sup_{P_1\in\mathcal{M}_1}\big(1-\hat\phi(P_1)\big)+a\cdot\sup_{P_0\in\mathcal{M}_0}\hat\phi(P_0)\ -\ a\cdot\alpha\ =\\ &\overset{(4.41)}{=}\ (1+a)\cdot R_\pi\big((\overline{P}_0,\overline{P}_1),\hat\sigma,W\big)\ -\ a\cdot\alpha\ \ge\\ &\overset{(*)}{\ge}\ \inf_{\sigma\in\mathcal{T}_0(\mathcal{X},\{0,1\})}(1+a)\cdot R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big)\ -\ a\cdot\alpha\ =\\ &\overset{(4.41)}{=}\ \inf_{\varphi\in\mathcal{T}_0}\left(\sup_{P_1\in\mathcal{M}_1}\big(1-\varphi(P_1)\big)+a\cdot\sup_{P_0\in\mathcal{M}_0}\varphi(P_0)\right)\ -\ a\cdot\alpha\ =\\ &\overset{(4.39)}{=}\ \inf_{\varphi\in\mathcal{T}_0(\alpha)}\ \sup_{P_1\in\mathcal{M}_1}1-\varphi(P_1)\end{aligned}$$

where $(*)$ follows from Proposition 4.5.

In order to prove that the infimum in (4.38) is attained, put $\Lambda_{P_0}:\ \mathcal{T}\to\mathbb{R}\,,\ \ \phi\mapsto\phi(P_0)\,$. According to Theorem 4.22, $\mathcal{T}(\alpha)$ is compact because

$$\mathcal{T}(\alpha)\ =\ \bigcap_{P_0\in\mathcal{M}_0}\Lambda_{P_0}^{-1}\big([0,\alpha]\big)$$

is closed in $\mathcal{T}$. Compactness of $\mathcal{T}(\alpha)$ and lower semicontinuity of

$$\mathcal{T}(\alpha)\ \to\ \mathbb{R}\,,\qquad \phi\ \mapsto\ \sup_{P_1\in\mathcal{M}_1}1-\phi(P_1)$$

imply that the infimum in (4.38) is attained by some generalized level-α-test; cf. e.g. (Denkowski et al., 2003, Theorem 1.3.11). □

As already stated in the previous subsection, Theorem 4.20 follows as an easy corollary now:

Proof of Theorem 4.20: Let $\tilde{\varphi}$ be an ordinary test which fulfills (4.27) and let π be the prior distribution from Theorem 4.23.

Then, $\tilde{\varphi}$ solves testing problem (4.24) if and only if it solves testing problem (4.30) according to Theorem 4.25. Next, $\tilde{\varphi}$ solves testing problem (4.30) if and only if its corresponding randomization $\tilde{\sigma}$ minimizes the Bayes risk

$$R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big) \;=\; \sum_{\theta\in\{0,1\}} \pi_\theta \cdot \sup_{P_\theta\in\mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]\;, \qquad \sigma \,\in\, \mathcal{T}\big(\mathcal{X},\{0,1\}\big)$$

(over all generalized randomizations) according to Theorem 4.23. Finally, it follows from Proposition 4.5 that $\tilde{\sigma}$ minimizes the Bayes risk

$$R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big) \;=\; \sum_{\theta\in\{0,1\}} \pi_\theta \cdot \sup_{P_\theta\in\mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]\;, \qquad \sigma \,\in\, \mathcal{T}\big(\mathcal{X},\{0,1\}\big)$$

(over all generalized randomizations) if and only if it minimizes the Bayes risk

$$R_\pi\big((\overline{P}_0,\overline{P}_1),\sigma,W\big) \;=\; \sum_{\theta\in\{0,1\}} \pi_\theta \cdot \sup_{P_\theta\in\mathcal{M}_\theta} \sigma(P_\theta)[W_\theta]\;, \qquad \sigma \,\in\, \mathcal{T}_0\big(\mathcal{X},\{0,1\}\big)$$

(over all ordinary randomizations). □

4.3.3 Least favorable pairs in hypothesis testing

As stated in Subsection 4.1.1, the publication of Huber and Strassen (1973) led to a lot of further research about least favorable pairs. However, a view years before this seminal paper, the existence of least favorable pairs has already been shown by Baumann (1968) in a more general setup. The testing problem in this more general setup mathematically coincides with the imprecise testing problem (4.23) in Subsection 4.3.1 where the hypotheses consist of coherent upper previsions. As a matter of fact, the following definition of least favorable pairs is only a reformulation of (Baumann, 1968, Definition 4.8) in terms of coherent upper previsions. In case of F-probabilities, this has already been done in (Augustin, 1998, § 3.3.2).

In the present subsection, let $\mathcal{X}$ be a set with algebra $\mathcal{A}$ and let $\alpha \in [0,1]$ be a fixed bound on the type I error. Recall that $\mathcal{T}_0$ and $\mathcal{T}$ denotes the set of all ordinary and generalized tests respectively.

Let $\overline{P}_0$ and $\overline{P}_1$ be coherent upper previsions on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ with credal sets $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively . In testing problem

$$\overline{P}_0 \qquad \text{vs.} \qquad \overline{P}_1$$

the infimal type II error over all level-α-tests is denoted by

$$\beta(\mathcal{M}_0,\mathcal{M}_1) \;:=\; \inf\Big\{ \sup_{P_1\in\mathcal{M}_1} P_1[1-\varphi] \;\Big|\; \varphi\in\mathcal{T}_0,\;\; P_0[\varphi]\le\alpha \;\;\forall\, P_0\in\mathcal{M}_0\Big\}$$

For any $P_0 \in \mathcal{M}_0$ and $P_1 \in \mathcal{M}_1$, the infimal type II error over all level-α-tests in testing problem

$$P_0 \qquad \text{vs.} \qquad P_1$$

is denoted by

$$\beta(P_0,P_1) \;:=\; \inf\big\{P_1[1-\varphi] \;\big|\; \varphi\in\mathcal{T}_0,\;\; P_0[\varphi]\le\alpha\big\}$$

Recall that, according to Theorem 4.25, it does not matter if these infima are calculated over all ordinary or generalized tests.

Definition 4.26 (Least favorable pairs) *Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be credal sets of coherent upper previsions on $(\Omega, \mathcal{A})$.*
$(\tilde{P}_0, \tilde{P}_1) \in \mathcal{M}_0 \times \mathcal{M}_1$ is a least favorable pair *if*

$$\beta(\mathcal{M}_0, \mathcal{M}_1) \;=\; \beta(\tilde{P}_0, \tilde{P}_1)$$

The following theorem essentially rephrases the main result in Baumann (1968) and is the analog to (Augustin, 1998, Satz 3.14), which is concerned with (continuous) F-probabilities.[7]

Theorem 4.27 *Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be credal sets of coherent upper previsions on $(\Omega, \mathcal{A})$. Then, there is a least favorable pair*

$$(\tilde{P}_0, \tilde{P}_1) \;\in\; \mathcal{M}_0 \times \mathcal{M}_1$$

Now, it is possible to prove this statement by use of the results of the previous subsections. Therefore, this proof is independent of Baumann (1968).

Proof: Let Θ, π and W be chosen as in Theorem 4.23. Then, $\pi_1 > 0$. Furthermore, $\pi = 0$ if and only if $a = 0$ where $a \in [0, \infty)$ comes from Lemma 4.24 where $\mathcal{T}_* = \mathcal{T}$.

Let $a = 0$. Then, for any $(P_0, P_1) \in \mathcal{M}_0 \times \mathcal{M}_1$,

$$0 \;\leq\; \beta(P_0, P_1) \;\leq\; \beta(\mathcal{M}_0, \mathcal{M}_1) \;\overset{(\dagger)}{=}\; \inf_{\phi \in \mathcal{T}} \sup_{P_1 \in \mathcal{M}_1} \big(1 - \phi(P_1)\big) \;=\; 0$$

where (†) follows from Lemma 4.24 for $\mathcal{T}_* = \mathcal{T}$.[8] That is, every (P_0, P_1) is a least favorable pair in case of $a = 0$.

Now, let $a > 0$; i.e., $\pi_\theta > 0 \;\; \forall \theta \in \{0, 1\}$.

[1] Firstly, note that

$$d_0 c_0 + d_1 c_1 \;=\; d_0 \overline{c}_0 + d_1 \overline{c}_1 \,, \qquad c_0 \leq \overline{c}_0 \,, \quad c_1 \leq \overline{c}_1$$

implies

$$c_0 \;=\; \overline{c}_0 \,, \qquad c_1 \;=\; \overline{c}_1$$

for $d_i \in (0, \infty)$, $c_i \in \mathbb{R}$, $\overline{c}_i \in \mathbb{R}$, $i \in \{0, 1\}$.

[2] According to the proof of Lemma 4.6 (a),

$$(P_0, P_1) \;\mapsto\; R_\pi\big((P_0, P_1), \sigma, W\big)$$

is continuous on the compact $\mathcal{M}_0 \times \mathcal{M}_1$ for every $\sigma \in \mathcal{T}_0(\mathcal{X}, \{0, 1\})$. So, (Denkowski et al., 2003, Theorem 1.3.11) implies the existence of some $(\tilde{P}_0, \tilde{P}_1) \in \mathcal{M}_0 \times \mathcal{M}_1$ such that

$$\begin{aligned} \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \sigma, W\big) \;&=\\ =\; \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} \; \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((P_0, P_1), \sigma, W\big) & \qquad (4.42) \end{aligned}$$

[7] In (Baumann, 1968, Korollar 5.6), the bound on the type I error may depend on $P_0 \in \mathcal{M}_0$ while the bound α is a constant here. Therefore, (Baumann, 1968, Korollar 5.6) is slightly more general than Theorem 4.27.

[8] Recall from Theorem 4.25 that it does not matter whether $\beta(\mathcal{M}_0, \mathcal{M}_1)$ is calculated over all generalized or ordinary level-α-tests.

According to Theorem 4.25, there is a generalized test $\tilde{\phi}' \in \mathcal{T}(\alpha)$ which achieves the infimum in (4.38). A suitable convex combination of $\tilde{\phi}'$ and

$$\phi_1 : \quad \mathrm{ba}(\Omega, \mathcal{A}) \to \mathbb{R}, \qquad \mu \mapsto \mu[I_\Omega]$$

yields some $\tilde{\phi} \in \mathcal{T}(\alpha)$ so that

$$\sup_{P_1 \in \mathcal{M}_1} 1 - \tilde{\phi}(P_1) \;\le\; \sup_{P_1 \in \mathcal{M}_1} 1 - \phi(P_1), \qquad \sup_{P_0 \in \mathcal{M}_0} \tilde{\phi}(P_0) \;=\; \alpha \tag{4.43}$$

That is, $\tilde{\phi}$ also achieves the infimum in (4.38) and, therefore

$$\sup_{P_1 \in \mathcal{M}_1} 1 - \tilde{\phi}(P_1) \;=\; \beta(\mathcal{M}_0, \mathcal{M}_1) \tag{4.44}$$

[3] Theorem 4.23 together with Proposition 4.5 implies

$$\begin{aligned} &\sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \tilde{\sigma}, W\big) \;= \\ &\qquad = \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} \; \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \sigma, W\big) \end{aligned} \tag{4.45}$$

where $\tilde{\sigma} \in \mathcal{T}(\mathcal{X}, \{0,1\})$ denotes the generalized randomization which corresponds to $\tilde{\phi}$. Next,

$$\begin{aligned} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \tilde{\sigma}, W\big) \;&\le\; \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \tilde{\sigma}, W\big) \;= \\ &\overset{(4.45)}{=} \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} \; \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \sigma, W\big) \;= \\ &\overset{\text{Lemma 4.6}}{=} \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} \; \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((P_0, P_1), \sigma, W\big) \;= \\ &\overset{(4.42)}{=} \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \sigma, W\big) \end{aligned}$$

Hence, according to Proposition 4.1,

$$R_\pi\big((\tilde{P}_0, \tilde{P}_1), \tilde{\sigma}, W\big) \;=\; \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \sigma, W\big) \tag{4.46}$$

and $\tilde{\sigma}$ is also optimal in the *decision problem* where the imprecise model $(\overline{P}_0, \overline{P}_1)$ is replaced by the precise model $(\tilde{P}_0, \tilde{P}_1)$. In the following part of the proof, it is shown, that its corresponding test $\tilde{\phi}$ also solves the *testing problem* where $(\overline{P}_0, \overline{P}_1)$ is replaced by $(\tilde{P}_0, \tilde{P}_1)$ – that is,

$$1 - \tilde{\phi}(\tilde{P}_1) \;=\; \beta\big(\tilde{P}_1, \tilde{P}_2\big)$$

[4] The definition of the Bayes risk and

$$\begin{aligned} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \tilde{\sigma}, W\big) \;&\overset{(4.46)}{=}\; \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((\tilde{P}_0, \tilde{P}_1), \sigma, W\big) \;= \\ &\overset{(4.42)}{=} \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} \; \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} R_\pi\big((P_0, P_1), \sigma, W\big) \;= \\ &\overset{\text{Lemma 4.6}}{=} \inf_{\sigma \in \mathcal{T}_0(\mathcal{X}, \{0,1\})} \; \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \sigma, W\big) \;= \\ &\overset{(4.45)}{=} \sup_{P_0 \in \mathcal{M}_0,\, P_1 \in \mathcal{M}_1} R_\pi\big((P_0, P_1), \tilde{\sigma}, W\big) \end{aligned} \tag{4.47}$$

imply

$$\pi_0\tilde{\sigma}(\tilde{P}_0)[W_0] + \pi_1\tilde{\sigma}(\tilde{P}_1)[W_1] \;=\; \pi_0 \sup_{P_0\in\mathcal{M}_0} \tilde{\sigma}(P_0)[W_0] + \pi_1 \sup_{P_1\in\mathcal{M}_1} \tilde{\sigma}(P_1)[W_1]$$

Hence,

$$\tilde{\phi}(\tilde{P}_0) \;\overset{(4.31)}{=}\; \tilde{\sigma}(\tilde{P}_0)[W_0] \;\overset{(\ddagger)}{=}\; \sup_{P_1\in\mathcal{M}_1} \tilde{\sigma}(P_1)[W_1] \;\overset{(4.43)}{=}\; \alpha \tag{4.48}$$

where (‡) follows from part [1] of the present proof.
Now, let $\varphi \in \mathcal{T}_0$ be any ordinary test which is not worse than $\tilde{\phi}$ in testing problem

$$\tilde{P}_0 \qquad \text{vs.} \qquad \tilde{P}_1$$

That is,

$$1-\varphi(\tilde{P}_1) \;\le\; 1-\tilde{\phi}(\tilde{P}_1)\,, \qquad \varphi(\tilde{P}_0) \;\le\; \alpha \tag{4.49}$$

Let $\sigma \in \mathcal{T}_0(\mathcal{X},\{0,1\})$ be the ordinary randomization which corresponds to the test φ. Then, it follows from (4.31), (4.32) (4.49) and (4.48) that

$$\sigma(\tilde{P}_1)[W_1] \;\le\; \tilde{\sigma}(\tilde{P}_1)[W_1]\,, \qquad \sigma(\tilde{P}_0)[W_0] \;\le\; \tilde{\sigma}(\tilde{P}_0)[W_0] \tag{4.50}$$

Next, the definition of the Bayes risk imply

$$\begin{aligned} R_\pi\big((\tilde{P}_0,\tilde{P}_1),\sigma,W\big) \;&=\; \pi_0\sigma(\tilde{P}_0)[W_0] + \pi_1\sigma(\tilde{P}_1)[W_1] \;\le \\ \overset{(4.50)}{\le}\;& \pi_0\tilde{\sigma}(\tilde{P}_0)[W_0] + \pi_1\tilde{\sigma}(\tilde{P}_1)[W_1] \;=\; R_\pi\big((\tilde{P}_0,\tilde{P}_1),\tilde{\sigma},W\big) \;\le \\ \overset{(4.46)}{\le}\;& R_\pi\big((\tilde{P}_0,\tilde{P}_1),\sigma,W\big) \end{aligned}$$

and, therefore,

$$\pi_0\sigma(\tilde{P}_0)[W_0] + \pi_1\sigma(\tilde{P}_1)[W_1] \;=\; \pi_0\tilde{\sigma}(\tilde{P}_0)[W_0] + \pi_1\tilde{\sigma}(\tilde{P}_1)[W_1]$$

Together with (4.50), this implies

$$\pi_1\sigma(\tilde{P}_1)[W_1] \;=\; \pi_1\tilde{\sigma}(\tilde{P}_1)[W_1]$$

according to part [1] of the present proof. That is,

$$1-\tilde{\phi}(\tilde{P}_1) \;=\; 1-\varphi(\tilde{P}_1)$$

and, therefore, we have shown that

$$1-\tilde{\phi}\big(\tilde{P}_1\big) \;=\; \beta\big(\tilde{P}_1,\tilde{P}_2\big) \tag{4.51}$$

[5] Finally,

$$\begin{aligned} \beta\big(\tilde{P}_1,\tilde{P}_2\big) \;&\overset{(4.51)}{=}\; 1-\tilde{\phi}\big(\tilde{P}_1\big) \;\ge\; \sup_{P_1\in\mathcal{M}_1} 1-\tilde{\phi}(P_1) \;\overset{(4.44)}{=}\; \beta(\mathcal{M}_1,\mathcal{M}_2) \;\ge \\ &\ge\; \beta\big(\tilde{P}_1,\tilde{P}_2\big) \end{aligned}$$

(where the last inequality is a trivial consequence of the definitions) implies that $\big(\tilde{P}_1,\tilde{P}_2\big)$ is a least favorable pair.

□

Chapter 5

Natural extensions and the sample space

5.1 Introduction

It has been detailedly pointed out in Subsection 3.4.1 that sample spaces are mathematical constructs and that the choice of a concrete sample space $(\mathcal{X}, \mathcal{A})$ in a given application is usually rather arbitrary. This is true in classical mathematical statistics and decision theory but it gets even more visible in case of imprecise probabilities. In order to see this, note that coherent upper previsions are functionals

$$\overline{P} \ : \ \mathcal{K} \ \to \ \mathbb{R}$$

which may be defined on any domain $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ where $(\mathcal{X}, \mathcal{A})$ is any sample space. However, in the mathematical evaluation, we are not tied up with the initial choice of $\mathcal{K}$. In fact, it is one of the most pleasant benefits of the use of imprecise probabilities that we can always extend $\overline{P}$ in a coherent way on larger domains[1] if this is convenient. This is possible by applying the method of natural extension developed in (Walley, 1991, § 3). The method of natural extension undoubtedly is one of the most important cornerstones of the whole theory of imprecise probabilities due to P. Walley – as stated in (Walley, 1991, p. 121):

> "After the ideas of avoiding sure loss and coherence (. . .), the most important concept in the present theory is that of natural extension. It is the fundamental concept in our theory of statistical inference (. . .)."

Accordingly, the method of natural extension is contained in every survey of the theory of coherent lower/upper previsions (e.g. Miranda (2008)). Due to its importance, the method of natural extension itself is still also a matter of resent research; see e.g. Pelessoni and Vicig (2005), Miranda and de Cooman (2007) and de Cooman et al. (2008).

However, the generous use of this method raises two questions which have not been considered so far. Answering both questions is highly appreciable since satisfactory answers to these questions are of great importance in oder to justify the use of natural extensions. The first question is concerned with extensions from $\mathcal{K}$ to the whole space $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$,

[1] As will be seen in Subsection 5.3.2, sometimes, it is even possible to restrict $\overline{P}$ on a smaller domain without loosing anything.

the second question is concerned with extensions of the sample space $(\mathcal{X},\mathcal{A})$ to a sample space $(\mathcal{X},\mathcal{A}')$ where $\mathcal{A}' \supset \mathcal{A}$:

Firstly, let

$$\overline{P} \; : \; \mathcal{K} \;\to\; \mathbb{R}$$

be a coherent upper prevision on any set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$. The reason for the use of coherent upper previsions instead of linear previsions or precise probabilities is the fact that it is far more realistic to give an upper bound on the previsions/expectations/probabilities than to precisely specify these quantities in applications. However, as a matter of fact, the upper prevision $\overline{P}[f]$ is again a precise real number and it is suggesting to state that this number $\overline{P}[f]$ is usually not precisely known in real applications. So, a practitioner will hardly be able to decide if $\overline{P}$ is the "correct" upper prevision or if another upper previson

$$\overline{P}' \; : \; \mathcal{K} \;\to\; \mathbb{R}$$

is the correct one where

$$\big|\overline{P}[f] - \overline{P}'[f]\big| \;<\; \varepsilon \qquad \forall\, f \in \mathcal{K}$$

for some (very) small $\varepsilon > 0$. As long as we only deal with elements f of $\mathcal{K}$, we may hope that this will only have small effects on the results. However, what happens if we apply the methods of natural extension in order to deal with functions $f \notin \mathcal{K}$? Is the natural extension of $\overline{P}'$ still close to the natural extension of $\overline{P}$? The investigations in Section 5.2 show that, unfortunately, the answer to this question is not affirmative. Even more, arbitrarily small changes in $\overline{P}$ on $\mathcal{K}$ can have arbitrarily large effects on its natural extension in general and, therefore, applying natural extensions may lead to meaningless results. An example where this happens in given in Subsection 5.2.1. Fortunately, not all is lost. In Subsection 5.2.2, it is shown that it can be guaranteed in many situation that small changes in $\overline{P}$ on $\mathcal{K}$ only have small effects on the natural extension. Though these results are not fully satisfactory, they show that it is possible to derive sensible conditions which protect from instable natural extensions. Hopefully, these initial investigations serve as a starting point for more sophisticated investigation into this direction.

Secondly, Section 5.3 is concerned with changes of the sample space: Again, let

$$\overline{P} \; : \; \mathcal{K} \;\to\; \mathbb{R}$$

be a coherent upper prevision on any $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$. Then, $\overline{P}$ can again be extended to a coherent upper prevision on the whole sample space $(\mathcal{X},\mathcal{A})$ so that we get a coherent upper prevision

$$\overline{P} \; : \; \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \;\to\; \mathbb{R}$$

However, we are not tied up with this choice of the sample space. Let $\mathcal{A}'$ be an algebra such that $\mathcal{A}' \supset \mathcal{A}$. Then, we can still extend $\overline{P}$ on the larger sample space $(\mathcal{X},\mathcal{A}')$ so that we get a coherent upper prevision

$$\overline{P} \; : \; \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \;\to\; \mathbb{R}$$

Therefore, it seems to be always possible to arbitrarily chose that algebra which is most convenient. However, at least in decision theory and, especially in statistics, choosing $\mathcal{A}$ or the larger $\mathcal{A}'$ has a fundamental effect on the evaluations. As can be seen from

the decision theoretic definitions in Chapter 3, the choice of the sample space determines the (randomized) decision functions which may be applied in the decision problem. In this way, extending the sample space leads to a larger set of valid (randomized) decision functions.
Let $(\overline{P}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{X},\mathcal{A})$, let W be a loss function and let $\tilde{\tau}$ be an optimal randomized decision function. That is, $\tilde{\tau}$ minimizes the Bayes risk $R\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big)$ over all randomized decision functions τ on $(\mathcal{X},\mathcal{A})$. By use of natural extensions, $(\overline{P}_\theta)_{\theta\in\Theta}$ turns into an imprecise model on $(\mathcal{X},\mathcal{A}')$. After that, $\tilde{\tau}$ still is a valid randomized decision function and it is easy to see that its Bayes risk is not affected by natural extension. However, the important question arises if optimality gets lost! This is because natural extension increases the set of all valid randomized decision functions and it is suggesting that there might be a randomized decision function on $(\mathcal{X},\mathcal{A}')$ which is better than $\tilde{\tau}$. In this case, natural extension would turn an optimal randomized decision function into a suboptimal one. Therefore, it seems to be most adequate to pose the following definition of optimality:

> *A randomized decision function is otimal if it minimizes the Bayes risk over all randomized decision function for any natural extension of the model.*

This definition is not very comfortable because it actually forces to always consider the whole power set of $\mathcal{X}$ – and the power set may be too large to be handled successfully. Especially in case of $\mathcal{X}=\mathbb{R}$, this would be very cumbersome.

Fortunately, the investigations in Section 5.3 show that such a proceeding is not necessary. In Subsection 5.3.1, it is proven that – after applying natural extension – there is no (randomized) decision function on $(\mathcal{X},\mathcal{A}')$ which is better than the best (randomized) decision function on $(\mathcal{X},\mathcal{A})$. The proof of this result turns out to be rather involved and heavily relies on the previous investigations in Chapter 3 and Chapter 4. Especially, it requires the general decision theoretic setup developed in Section 3.3 on base of L. Le Cam's work, results from Section 4.2 based on topological properties and an application of the theory of vector lattices.
Next, it is described in Subsection 5.3.2 how the result can also be applied the other way round: Sometimes it enables to reduce the sample space without loosing anything because the optimal (randomized) decision function is guaranteed to live on a smaller sample space.

Chapter 5 closes with an application in Section 5.4 which is based on the answers to both above questions given in Section 5.2 and Section 5.3. This application lies in discretizing – a topic which increasingly attracts attention within the theory of imprecise probabilities; cf. Obermeier and Augustin (2007) and Troffaes (2008). In short, this is done as follows: Let $(\overline{P}'_\theta)_{\theta\in\Theta}$ be an imprecise model such that each

$$\overline{P}'_\theta \;:\; \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \;\to\; \mathbb{R}$$

is the natural extension of a coherent upper prevision on a finite subset $\mathcal{K}\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. Firstly, each element of $\mathcal{K}$ is discretized such that we get a corresponding set $\hat{\mathcal{K}}$ of discrete functions which are close to the elements of $\mathcal{K}$. Next, it is possible to use $\hat{\mathcal{K}}$ in order to define an imprecise model on a discrete sample space $(\mathcal{X},\mathcal{A})$ which is – according to Section 5.2 – close to the original imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$. Next, it can be shown that – according to Section 5.3 – the solution of the discretized decision problem on $(\mathcal{X},\mathcal{A})$ is an approximate solution of the original decision problem on $(\mathcal{X},\mathcal{A}')$.

5.2 Instability of the natural extension

5.2.1 A first example

Usually, the sole use of a simple parametric model consisting of precise probabilities is hardly justifiable in real applications because real data almost never stem from such a model and, if they would do so, we could not be sure that they really do. It is a well known fact that small deviations from a precise model can have large effects on the statistical methods – cf. e.g. Huber (1981). However, a precise model is usually not precisely true. This is one reason for the use of imprecise probabilities. Of course, it is far more easy to determine upper and lower bounds for the probabilities than to determine precise probabilities. Tough it will not be possible to precisely determine correct upper and lower bounds, small changes in the upper and lower bounds should only have small effects in the statistical evaluation. This is usually true but, unfortunately, this is not always true. Arbitrarily small changes in the upper and lower bounds can have arbitrarily large effects in some cases. This is because the theory of imprecise probabilities commonly uses a method which is potentially most instable – namely natural extension.

Especially in applications, the method of natural extension is a frequently used comfortable tool because it enables to define a coherent upper prevision $\overline{P}$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ or on a subset of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ in the following way: An experimenter determines an upper prevision

$$\overline{P} : \ \mathcal{K} \ \rightarrow \ \mathbb{R}$$

on some subset $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ and then extends this prevision to a coherent lower prevision on a larger set – by means of a natural extension, cf. Section 2.3. For simplicity of notation, we may assume that he extends the prevision to the whole set $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$.

In general, such a proceeding can be very instable and may lead to arbitrary results. To see this, let us consider the following simple example:

Put $\mathcal{X} = [0,1]$ and let $\mathcal{A}$ be the Borel-σ-algebra of $[0,1]$. Put

$$f_0 : \ [0,1] \ \rightarrow \ \mathbb{R}\,, \qquad x \ \mapsto \ x$$

and $\mathcal{K} = \{f_0\}$. Furthermore,

$$\overline{P}[f_0] = 0$$

and

$$\overline{P}'[f_0] = \varepsilon$$

where $0 < \varepsilon < 1$. Then, the natural extensions are given by

$$\overline{P}[f] \ = \ \sup_{P \in \mathcal{M}} P[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

and

$$\overline{P}'[f] \ = \ \sup_{P' \in \mathcal{M}'} P'[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

where

$$\mathcal{M} \ = \ \big\{P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}) \ \big| \ P[f_0] = 0\big\}$$
$$\mathcal{M}' \ = \ \big\{P' \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}) \ \big| \ P'[f_0] \in [0,\varepsilon]\big\}$$

For every $P \in \mathcal{M}$, it follows from $f_0 \;\geq\; \varepsilon I_{[\varepsilon,1]}$ that

$$0 \;=\; P[f_0] \;\geq\; P\big[\varepsilon I_{[\varepsilon,1]}\big] \;=\; \varepsilon\, P\big[I_{[\varepsilon,1]}\big] \;\geq\; 0 \tag{5.1}$$

and therefore, $P\big[I_{[\varepsilon,1]}\big] \;=\; 0$. Hence,

$$\overline{P}\big[I_{[\varepsilon,1]}\big] \;=\; 0$$

Let δ_ε be the Dirac measure in ε. Then $\delta_\varepsilon[f_0] = \varepsilon$ implies $\delta_\varepsilon \;\in\; \mathcal{M}'$ and

$$1 \;=\; \sup_{x\in[0,1]} I_{[\varepsilon,1]}(x) \;\geq\; \sup_{P'\in\mathcal{M}'} P'\big[I_{[\varepsilon,1]}\big] \;\geq\; \delta_\varepsilon\big[I_{[\varepsilon,1]}\big] \;=\; 1$$

Hence,

$$\overline{P}'\big[I_{[\varepsilon,1]}\big] \;=\; 1$$

Summing up, we have

$$\overline{P}\big[I_{[\varepsilon,1]}\big] \;=\; \inf_{x\in[0,1]} I_{[\varepsilon,1]}(x)\,, \qquad \overline{P}'\big[I_{[\varepsilon,1]}\big] \;=\; \sup_{x\in[0,1]} I_{[\varepsilon,1]}(x)$$

This is, indeed, the worst thing that can happen. The unpleasant message of this example is:

> *Determining a coherent upper prevision on some functions $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ in a first step and extending the coherent upper prevision (by natural extension) to some functions $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ in a second step may lead to arbitrary results: Arbitrarily small changes of one upper bound*
>
> $$\overline{P}[f_0]$$
>
> *(where $f_0 \in \mathcal{K}$) may have arbitrarily large effects on the bounds*
>
> $$\overline{P}[f]\,, \qquad f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

Note that the above example is not a pathological one: The sample space is a compact interval in $\mathbb{R}$, the algebra $\mathcal{A}$ is the Borel-σ-algebra and the coherent upper prevision on $\mathcal{K}$ is a very easy one because $\mathcal{K}$ only consists of one element f_0 and this f_0 is a linear function. It would even have made no difference if we would have taken $\mathcal{K}$ to be the linear space

$$\mathcal{K} \;:=\; \{a f_0 \mid a \in \mathbb{R}\}$$

However, the above example, indeed, is somehow special because we have

$$\underline{P}[f_0] = \overline{P}[f_0]$$

and this is a precise prevision which is not really what we want in imprecise probabilities. Nevertheless, the use of such imprecise probabilities where

$$\underline{P}[f] \;=\; \overline{P}[f] \tag{5.2}$$

at least for some (non-constant) functions $f \in \mathcal{K}$ is not unusual in applications of imprecise probabilities. Though such precise values (5.2) of imprecise probabilities seem to be problematic, this problem does not arise in classical probability theory (where upper and lower bounds always coincide) because, there, it is not possible to change single values in the above manner and something like a method of natural extension does not exist anyway.

One might argue that the coherent upper prevision used in the above example is not a good one and that it is enough to take a short look at it in order to see this. However, this example is also an extremely simple one and it is hard to guarantee that such "detecting bad models by a short look at it" still works for more complicated previsions.

The above example does not show that using natural extensions was indefensible in real applications but it shows that natural extension should not be used unthoughtfully.

For applications, it would be desirable to have some guidelines which prevent practitioners from arbitrary results because of an instable natural extension. The following subsection makes a first attempt in this direction but it certainly does not succeed in giving a final, satisfactory answer. Hopefully, future research will provide some more insight into this topic.

5.2.2 Stable imprecise probabilities

The present subsection is concerned with conditions that protect against instable natural extensions. The example presented in Subsection 5.2.1 shows that it does not seem to be promising to solely put restrictions on $\mathcal{K}$ such as "$\mathcal{K}$ should contain only a small number of functions", "$\mathcal{K}$ should be a linear space", "the functions in $\mathcal{K}$ should be continuous/monotone/linear" or something like that.
Instead, it can be seen from Equation (5.1) that Condition

$$\underline{P}[f_0] \;=\; \overline{P}[f_0] \tag{5.3}$$

is in fact crucial for the instability of the natural extension in the example presented in Subsection 5.2.1. Therefore, it is suggesting to avoid instability of the natural extension by avoiding (5.3). This is done in the following proposition:

Proposition 5.1 *Let $\overline{P}$ be a coherent upper prevision on a set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ and let*

$$\mathcal{F} \;:=\; \big\{f_1\,,\,\dots\,,\,f_n\big\} \;\subset\; \mathcal{K}$$

be a finite subset of $\mathcal{K}$.
Let $\overline{P}'$ be another coherent upper prevision on $\mathcal{K}$ such that

$$\overline{P}'[f] \;=\; \overline{P}[f] \qquad \forall\, f \in \mathcal{K} \setminus \mathcal{F}$$

and, for some real numbers $0 < \varepsilon_i < 1\,,\; i \in \{1,\dots,n\}\,,$

$$\overline{P}[f_i] \;\le\; \overline{P}'[f_i] \;\le\; \overline{P}[f_i] + \varepsilon_i\big(\overline{P}[f_i] - \underline{P}[f_i]\big) \qquad \forall\, i \in \{1,\dots,n\} \tag{5.4}$$

where $\underline{P}$ is the coherent lower prevision on $\mathcal{K}$ which corresponds to $\overline{P}$. [2] *Let $\underline{P}$, $\overline{P}$ and $\overline{P}'$ also denote the respective natural extensions of $\underline{P}$, $\overline{P}$ and $\overline{P}'$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$.*
Then, for $\epsilon := \varepsilon_1 + \dots + \varepsilon_n$:

[2] $\underline{P}[f] = \inf_{P\in\mathcal{M}} P[f] \;\; \forall\, f \in \mathcal{K}$ where $\mathcal{M}$ is the credal set of $\overline{P}$.

a) $\overline{P}[f] \ \leq \ \overline{P}'[f] \ \leq \ \overline{P}[f] + \varepsilon\big(\sup f - \underline{P}[f]\big) \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$

b) If $\varepsilon < 1$,

$$\overline{P}[f] \ \leq \ \overline{P}'[f] \ \leq \ \overline{P}[f] + \frac{\varepsilon}{1-\varepsilon}\cdot\big(\overline{P}[f] - \underline{P}[f]\big) \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

That is, Proposition 5.1 investigates what may happen (or rather what cannot happen) if some values of a coherent upper prevision $\overline{P}$ are made slightly larger. "Slightly" means: a small percentage of $\overline{P}[f] - \underline{P}[f]$. Therefore, $\overline{P}[f]$ may not be changed if $\overline{P}[f] = \underline{P}[f]$ so that the example in Subsection 5.2.1 is excluded. So, Proposition 5.1 really explains how to avoid instability of the natural extension by avoiding such bottlenecks (5.3). Nevertheless, modeler will often create instable coherent upper previsions $\overline{P}_0$ where

$$\overline{P}_0[f_i] - \underline{P}_0[f_i]$$

is very small or equal to zero for some $f_i \in \mathcal{K}$. Then, it will often be sensible to turn this potentially most instable coherent upper prevision into a stable one – not as part of the decision theoretic evaluation but as part of modeling. A canonical and easy way to do this is as follows: Take any appropriately small $\alpha \in (0,1)$ and use

$$\overline{P}_\alpha \ : \ \mathcal{K} \ \to \ \mathbb{R}\,, \qquad f \ \mapsto \ (1-\alpha)\overline{P}_0[f] + \alpha \sup f \tag{5.5}$$

instead of $\overline{P}_0$. It is easy to see that $\overline{P}_\alpha$ is again a coherent upper prevision. Furthermore, we have

$$\begin{aligned} \overline{P}_\alpha[f] - \underline{P}_\alpha[f] \ &= \ (1-\alpha)\overline{P}_0[f] + \alpha \sup f \ - \ (1-\alpha)\underline{P}_0[f] - \alpha \inf f \ \geq \\ &\geq \ \alpha(\sup f - \inf f) \end{aligned}$$

for every $f \in \mathcal{K}$. Of course, the above example shows that going over to $\overline{P}_\alpha$ can massively change the results. But, if this happens, it is so much the better to go over to $\overline{P}_\alpha$ because it is usually impossible to guaranty that the "true" coherent upper prevision is certainly equal to $\overline{P}_0$ and does not lie somewhere between $\overline{P}_0$ and $\overline{P}_\alpha$. Furthermore, Proposition 5.1 implies that additional small enlargements of $\overline{P}_\alpha$ will only moderately change the results.

Proof of Proposition 5.1: Let $\mathcal{M}$ denote the credal set of $\overline{P}$ and $\mathcal{M}'$ denote the credal set of $\overline{P}'$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$. Note that

$$\mathcal{M} \ \subset \ \mathcal{M}' \qquad \text{hence} \qquad \overline{P}'[f] \ \geq \ \underline{P}[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \tag{5.6}$$

[1] Fix any $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ and any $P' \in \mathcal{M}'$. For every $i \in \{1,\dots,n\}$, there is a $P_i \in \mathcal{M}$ such that $\underline{P}[f_i] = P_i[f_i]$ (cf. Corollary 2.16).

- Put $P'_0 := P'$ and consider the following inductive definitions for $i \in \{1,\dots,n\}$:
  – In case of $P'_{i-1}[f_i] \leq \overline{P}[f_i]$ (CASE 1), put $\alpha_i = 0$.
  – In case of $P'_{i-1}[f_i] > \overline{P}[f_i]$ (CASE 2), put

$$\alpha_i \ := \ \frac{P'_{i-1}[f_i] - \overline{P}[f_i]}{P'_{i-1}[f_i] - P_i[f_i]}$$

Then, put

$$P'_i \ := \ (1-\alpha_i)P'_{i-1} \ + \ \alpha_i P_i \tag{5.7}$$

- By induction, we proof in the following that, for every $i \in \{0, \ldots, n\}$,

$$P'_i \;\in\; \mathcal{M}' \tag{5.8}$$

$$P'_i[f_j] \;\leq\; \overline{P}[f_j] \qquad \forall\, j \in \{1, \ldots, i\} \tag{5.9}$$

and

$$P'[f] \;\leq\; P'_i[f] + \sum_{j=1}^{i} \varepsilon_j \cdot \big(\overline{P}'[f] - \underline{P}[f]\big) \tag{5.10}$$

Obviously, (5.8), (5.9) and (5.10) are fulfilled for $i = 0$.
Next, let (5.8), (5.9) and (5.10) be fulfilled for $i - 1$.

  - CASE 1: In case of $P'_{i-1}[f_i] \leq \overline{P}[f_i]$, we have $\alpha_i = 0$, $P'_i = P'_{i-1}$ and, therefore, (5.8) and (5.9) are fulfilled. In addition, (5.10) follows because the induction hypothesis implies

$$\begin{aligned} P'[f] \quad &\leq \quad P'_{i-1}[f] + \sum_{j=1}^{i-1} \varepsilon_j \cdot \big(\overline{P}'[f] - \underline{P}[f]\big) \;\leq \\ &\overset{(5.6)}{\leq} \quad P'_i[f] + \sum_{j=1}^{i} \varepsilon_j \cdot \big(\overline{P}'[f] - \underline{P}[f]\big) \end{aligned}$$

  - CASE 2: In case of $P'_{i-1}[f_i] > \overline{P}[f_i]$, it follows from

$$\underline{P}[f_i] \;=\; P_i[f_i] \;\leq\; \overline{P}[f_i] \;<\; P'_{i-1}[f_i] \;\overset{(5.8)}{\leq}\; \overline{P}[f_i] + \epsilon_i\big(\,\overline{P}[f_i] - \underline{P}[f_i]\,\big)$$

that

$$0 \;\leq\; \alpha_i \;=\; \frac{P'_{i-1}[f_i] - \overline{P}[f_i]}{P'_{i-1}[f_i] - P_i[f_i]} \;\leq\; \frac{\varepsilon_i(\overline{P}[f_i] - \underline{P}[f_i])}{\overline{P}[f_i] - \underline{P}[f_i]} \;=\; \varepsilon_i \tag{5.11}$$

then. Especially, $\alpha_i \in [0, 1]$. Next, the definition of P'_i, (5.6) and the induction hypothesis immediately imply the validity of (5.8) for i and

$$P'_i[f_j] \;\leq\; \overline{P}[f_j] \quad \forall\, j \in \{1, \ldots, i-1\}$$

Furthermore,

$$\begin{aligned} P'_i[f_i] \quad &= \quad (1-\alpha_i)P'_{i-1}[f_i] + \alpha P_i[f_i] \;= \\ &= \quad \frac{\overline{P}[f_i] - P_i[f_i]}{P'_{i-1}[f_i] - P_i[f_i]} \cdot P'_{i-1}[f_i] \;+\; \frac{P'_{i-1}[f_i] - \overline{P}[f_i]}{P'_{i-1}[f_i] - P_i[f_i]} \cdot P_i[f_i] \;= \\ &= \quad \overline{P}[f_i] \end{aligned}$$

That is, we have proven the validity of (5.8) and (5.9) for i so far. In order to prove (5.10), note that

$$\begin{aligned} P'_{i-1}[f] \quad &= \quad \alpha_i P'_{i-1}[f] + (1-\alpha_i)P'_{i-1}[f] + \alpha_i P_i[f] - \alpha_i P_i[f] \;= \\ &\overset{(5.7)}{=} \quad \alpha_i P'_{i-1}[f] + P'_i[f] - \alpha_i P_i[f] \;\leq \\ &\overset{(5.8)}{\leq} \quad P'_i[f] + \alpha_i\big(\overline{P}'[f] - \underline{P}[f]\big) \;\leq \\ &\leq \quad P'_i[f] + \varepsilon_i\big(\overline{P}'[f] - \underline{P}[f]\big) \end{aligned}$$

where the last inequality follows from (5.6) and (5.11). Together with the induction hypothesis, this implies

$$\begin{aligned} P'[f] \ &\leq\ P'_{i-1}[f] + \sum_{j=1}^{i-1} \varepsilon_j \cdot \bigl(\overline{P}'[f] - \underline{P}[f]\bigr) \ \leq \\ &\leq\ P'_i[f] + \sum_{j=1}^{i} \varepsilon_j \cdot \bigl(\overline{P}'[f] - \underline{P}[f]\bigr) \end{aligned}$$

Summing up, we have proven by induction the validity of (5.8), (5.9) and (5.10) for every $i \in \{1, \dots, n\}$ so far.

- For $i = n$, (5.8) and (5.9) imply

$$P'_n \ \in\ \mathcal{M}$$

Hence, it follows from (5.10) and the definition of ε that

$$P'[f] \ \leq\ \overline{P}[f] + \varepsilon \cdot \bigl(\overline{P}'[f] - \underline{P}[f]\bigr) \tag{5.12}$$

[2] It is shown in part [1] of the present proof that (5.12) is valid for every $P' \in \mathcal{M}'$ and every $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$. Hence, we have

$$\overline{P}'[f] \ \leq\ \overline{P}[f] + \varepsilon \cdot \bigl(\overline{P}'[f] - \underline{P}[f]\bigr) \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}) \tag{5.13}$$

[3] Now, part a) of Proposition 5.1 follows from

$$\begin{aligned} \overline{P}'[f] \ &\overset{(5.13)}{\leq}\ \overline{P}[f] + \varepsilon \cdot \bigl(\overline{P}'[f] - \underline{P}[f]\bigr) \ \leq \\ &\leq\ \overline{P}[f] + \varepsilon \cdot \bigl(\sup f - \underline{P}[f]\bigr) \qquad\qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}) \end{aligned}$$

In case of $\varepsilon = \sum_{j=1}^{n} \varepsilon_j \ < \ 1$ a simple transformation of (5.13) yields

$$\overline{P}'[f] \ \leq\ \overline{P}[f] + \frac{\varepsilon}{1-\varepsilon} \cdot \bigl(\overline{P}[f] - \underline{P}[f]\bigr) \qquad\qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$$

that is Proposition 5.1 b).

□

Remark 5.2

a) Of course, Proposition 5.1 a) can be simplified to the weaker bound

$$\overline{P}[f] \ \leq\ \overline{P}'[f] \ \leq\ \overline{P}[f] + \varepsilon\bigl(\sup f - \inf f\bigr) \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$$

b) Proposition 5.1 is only concerned with the case where the coherent upper prevision is made slightly larger. This is because Proposition 5.1 has been derived especially for applications in discretizing and, for this purpose, Proposition 5.1 is indeed suitable; cf. Section 5.4. Nevertheless, Proposition 5.1 can also be used in order to get a similar version in that case where the coherent upper prevision is made slightly smaller. This may simply be done by interchanging the roles of $\overline{P}$ *and* $\overline{P}'$*:*

Now, let $\overline{P}'$ be a coherent upper prevision on $\mathcal{K}$ such that

$$\overline{P}'[f] \;=\; \overline{P}[f] \qquad \forall\, f \in \mathcal{K} \setminus \mathcal{F}$$

and, for every $i \in \{1, \ldots, n\}$,

$$\overline{P}[f_i] \;\geq\; \overline{P}'[f_i] \;\geq\; \overline{P}[f_i] - \varepsilon_i\big(\overline{P}'[f_i] - \underline{P}'[f_i]\big) \tag{5.14}$$

Then, for every $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$

$$\overline{P}[f] \;\geq\; \overline{P}'[f] \;\geq\; \overline{P}[f] - \varepsilon\big(\,\sup f - \inf f\,\big) \tag{5.15}$$

However, this version is not at all as good as the original one because the bound in (5.14) is based on

$$\overline{P}'[f] - \underline{P}'[f]$$

instead of

$$\overline{P}[f] - \underline{P}[f]$$

Of course, $\overline{P}'[f]$ is close to $\overline{P}[f]$ by assumption but it is not yet assured that $\underline{P}'[f]$ is close to $\underline{P}[f]$. If $\overline{P}'[f] - \underline{P}'[f]$ happens to be small, then this version of Proposition 5.1 is useless.

As pointed out in Remark 5.2, the above proposition is only concerned with the case where the coherent upper prevision is made slightly larger. Though this is enough for applications in discretizing[3], Proposition 5.1 can only serve as a starting point for more sophisticated examinations of imprecise probabilities which avoid instable natural extensions.

It would be most desirable to derive a result of the following form:

> Let $\overline{P}_0$ and $\overline{P}_0'$ be coherent upper previsions on a set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$. Then, for every $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$,
>
> $$b_1(\overline{P}_0, \overline{P}_0', f)) \;\leq\; \overline{P}[f] - \overline{P}'[f] \;\leq\; b_2(\overline{P}_0, \overline{P}_0', f)) \tag{5.16}$$
>
> where $\overline{P}$ and $\overline{P}'$ denote the natural extensions of $\overline{P}_0$ and $\overline{P}_0'$ respectively – and b_1 and b_2 provide some useful nontrivial bounds.

After talking to Matthias Troffaes about this problem at the *Fifth International Symposium on Imprecise Probability: Theories and Applications* in 2007, he gave me the hint that such problems have already been treated in linear programming. The above problem can indeed be formulated in terms of linear programming or – more general – in terms of convex optimization. Since the present subsection deals with (possibly) infinite spaces and linear programming may only be applied in case of finite spaces, more general convex optimization is considered now:

Again, let $\overline{P}_0$ and $\overline{P}_0'$ be coherent upper previsions on a set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$; let $\overline{P}$ and $\overline{P}'$ denote the respective natural extensions. Then, for every $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$, $\overline{P}[f]$ is the optimal value of the convex optimization problem

$$\sup_P P[f]\,, \qquad P \in \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A})\,, \qquad G(P) \leq 0 \tag{5.17}$$

[3] cf. Section 5.4

where G is the function

$$G: \ \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}) \ \rightarrow \ \left\{ b:\mathcal{K}\to\mathbb{R} \ \middle| \ \sup_{f_0\in\mathcal{K}} \frac{|b(f_0)|}{\|f_0\|} \ < \ \infty \right\}$$

which is given by

$$G(P)[f_0] = P[f_0] - \overline{P}_0[f_0] \qquad \forall\, P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}) \quad \forall\, f_0 \in \mathcal{K}$$

Accordingly, $\overline{P}'[f]$ is the optimal value of a convex optimization problem

$$\sup_P P[f]\,, \qquad P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A})\,, \qquad G'(P) \leq 0 \tag{5.18}$$

Now, we can formulate our problem in terms of convex optimization:

> In how far are the optimal values in (5.17) and (5.18) similar if $\overline{P}_0$ and $\overline{P}_0'$ are similar?

Answers to this problem are given by the so-called “sensitivity analysis” in convex optimization – cf. e.g. Luenberger (1969) for general convex optimization and Chvátal (1983) or Jansen et al. (1997) for linear programming.

Using one of the main results of sensitivity analysis in convex optimization[4], we get the following very general theorem:

Theorem 5.3 *Let $\overline{P}_0$ and $\overline{P}_0'$ be coherent upper previsions on a set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$. For a fixed $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$, let $\wp_f$ be a Lagrange multiplier of the convex optimization Problem (5.17) and let $\wp_f'$ be a Lagrange multiplier of the convex optimization Problem (5.18). $\overline{P}$ and $\overline{P}'$ denote the respective natural extensions of $\overline{P}_0$ and $\overline{P}_0'$. Then,*

$$\wp_f'\big(\overline{P}_0 - \overline{P}_0'\big) \ \leq \ \overline{P}[f] - \overline{P}'[f] \ \leq \ \wp_f\big(\overline{P}_0 - \overline{P}_0'\big)$$

Proof: This is a direct consequence of (Luenberger, 1969, §8.4). □

Note that such Lagrange multipliers need not exist. However, Lagrange multipliers exist under the following assumption; cf. e.g. (Luenberger, 1969, §8.3, Theorem 1):

> Assume that there are $\hat{P}\,,\ \hat{P}' \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}))$ and $\varepsilon,\ \varepsilon' > 0$ such that
>
> $$\hat{P}[f_0] + \varepsilon \ \leq \ \overline{P}_0[f_0] \qquad \text{and} \qquad \hat{P}'[f_0] + \varepsilon' \ \leq \ \overline{P}_0'[f_0] \tag{5.19}$$
>
> for every $f_0 \in \mathcal{K}$.

Apparently, this assumption is very similar to the assumptions in Proposition 5.1. It is always fulfilled if $\mathcal{K}$ is finite and

$$\underline{P}_0[f_0] < \overline{P}_0[f_0]\,, \qquad \underline{P}_0'[f_0] < \overline{P}_0'[f_0] \qquad \forall\, f_0 \in \mathcal{K} \tag{5.20}$$

This is the content of the following lemma:

Lemma 5.4 *Let $\overline{P}_0$ and $\overline{P}_0'$ be coherent upper previsions on a finite set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ such that (5.20) is fulfilled. Then, condition (5.19) is fulfilled.*

Proof: Let $\mathcal{M}$ be the credal set of $\overline{P}_0$. For every $f_i \in \mathcal{K} = \{f_1,\dots,f_n\}$, there is a $P_i \in \mathcal{M}$ such that $P_i[f_i] = \underline{P}_0[f_i]$. Then, put

$$\hat{P} := \frac{1}{n}\sum_{j=1}^{n} P_j \qquad \text{and} \qquad \varepsilon = \min_{i=1,\dots,n} \big(\overline{P}_0[f_i] - \hat{P}[f_i]\big)$$

and (5.20) guaranties $\varepsilon > 0$.
The same proof applies for $\overline{P}_0'$. □

[4] (Luenberger, 1969, §8.4)

5.3 Extensions and reductions of the sample space

5.3.1 Extension of Algebras and Natural Extension of Previsions

5.3.1.1 Description of the problem

As explained in the introductory Section 5.1, the arbitrariness of the choice of the sample space becomes apparent in particular in the theory of imprecise probabilities. There, the sample space may always be extended by the method of natural extension but, e.g. in case of decision theory, it is not obvious if such extensions does not change the results in a rather arbitrary way:

Let $\mathcal{X}$ be a set with algebras $\mathcal{A}$, $\mathcal{A}'$ such that that

$$\mathcal{A} \subset \mathcal{A}'$$

Let $\overline{P}$ be a coherent upper prevision on $\mathcal{X}$ with domain $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$:

$$\overline{P}: \quad \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \quad \longrightarrow \quad \mathbb{R}$$

By the method of natural extension, $\overline{P}$ may be extended to a coherent upper prevision on $\mathcal{X}$ with domain $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$:

$$\overline{P}': \quad \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \quad \longrightarrow \quad \mathbb{R}$$

The credal set of $\overline{P}$ on $(\mathcal{X},\mathcal{A})$ is denoted by

$$\mathcal{M} \;=\; \big\{P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}) \;\big|\; P[f] \;\le\; \overline{P}[f] \quad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})\big\} \tag{5.21}$$

and the credal set of $\overline{P}'$ on $(\mathcal{X},\mathcal{A}')$ is denoted by

$$\mathcal{M}' \;=\; \big\{P' \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \;\big|\; P'[f] \;\le\; \overline{P}'[f] \quad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})\big\} \tag{5.22}$$

Since such a natural extension of coherent upper previsions is always possible, it is stated in many articles that

> "*without loss of generality, we may assume that $\overline{P}$ is a coherent upper prevision on the larger domain $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$*"

where $\mathcal{A}'$ commonly happens to be the power set of $\mathcal{X}$. However, at least in decision theory, it changes a lot if the sample space is $(\mathcal{X},\mathcal{A})$ or $(\mathcal{X},\mathcal{A}')$ because the choice of the sample space determines the valid (randomized) decision functions.

Let $(\mathbb{D},\mathcal{D})$ be a decision space. In case of the sample space $(\mathcal{X},\mathcal{A})$, the valid randomized decision functions are given by the finitely additive Markov kernels

$$\tau \;:\; \mathcal{X}\times\mathcal{D} \;\to\; \mathbb{R}\,, \qquad (x,D) \;\mapsto\; \tau_x(D)$$

where the map $\tau_\bullet(D) \,:\, x \mapsto \tau_x(D)$ is assumed to be an element of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ for every $D \in \mathcal{D}$.

However, in case of the larger sample space $(\mathcal{X},\mathcal{A}')$, the valid randomized decision functions are given by the finitely additive Markov kernels

$$\tau' \;:\; \mathcal{X}\times\mathcal{D} \;\to\; \mathbb{R}\,, \qquad (x,D) \;\mapsto\; \tau'_x(D)$$

where the map $\tau'_\bullet(D) \,:\, x \mapsto \tau'_x(D)$ is only assumed to be an element of the larger set $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ for every $D \in \mathcal{D}$. That is: Extending the sample space leads to a larger set of valid (randomized) decision functions.

Now, let $(\overline{P}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{X},\mathcal{A})$, let

$$W \;:\; \Theta \times \mathbb{D} \;\to\; \mathbb{R}\,, \qquad (\theta,t) \;\mapsto\; W_\theta(t)$$

be a loss function and let $\tilde{\tau}$ be an optimal randomized decision function. That is[5], in case of the sample space $(\mathcal{X},\mathcal{A})$, $\tilde{\tau}$ minimizes the Bayes risk $R\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big)$ over all randomized decision functions τ such that $\tau_\bullet \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$. By use of natural extensions, $(\overline{P}_\theta)_{\theta\in\Theta}$ turns into an imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A}')$. After that, $\tilde{\tau}$ still is a valid randomized decision function and it is easy to see that its Bayes risk is not affected by natural extension – i.e.

$$R\big((\overline{P}'_\theta)_{\theta\in\Theta},\hat{\tau},W\big) \;=\; R\big((\overline{P}_\theta)_{\theta\in\Theta},\hat{\tau},W\big)$$

However, the important question raises if optimality gets lost. This is because natural extension increases the set of all valid randomized decision functions: Every finitely additive Markov kernel τ' such that $\tau'_\bullet \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ is a valid randomized decision function now and it is suggesting that there might be a randomized decision function on $(\mathcal{X},\mathcal{A}')$ such that

$$R\big((\overline{P}'_\theta)_{\theta\in\Theta},\tau',W\big) \;<\; R\big((\overline{P}'_\theta)_{\theta\in\Theta},\hat{\tau},W\big) \tag{5.23}$$

In this case, natural extension would turn an optimal randomized decision function into a suboptimal one. Since it is one of the fundamental properties of imprecise probabilities to extend the sample space whenever convenient, it seems to be most adequate to pose the following definition of optimality:

> *A randomized decision function is optimal if it minimizes the Bayes risk over all randomized decision function for any natural extension of the model.*

This definition is not very comfortable because it actually forces to always consider the whole power set of $\mathcal{X}$ – and the power set may be too large to be handled successfully. Especially in case of $\mathcal{X} = \mathbb{R}$, this would be very cumbersome.

Fortunately, such a proceeding is not necessary! This is shown by the main theorem, Theorem 5.6, of the present subsection. It states that (5.23) does not happen:

> *If $\tilde{\tau}$ minimizes the Bayes risk $R\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big)$ over all randomized decision functions τ such that $\tau_\bullet \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$, then there is no natural extension $(\overline{P}'_\theta)_{\theta\in\Theta}$ of the model and no randomized decision function τ' on the larger sample space such that*
>
> $$R\big((\overline{P}'_\theta)_{\theta\in\Theta},\tau',W\big) \;<\; R\big((\overline{P}'_\theta)_{\theta\in\Theta},\hat{\tau},W\big)$$

A mathematical rigorous formulation of this theorem is contained in the following subsection. Its proof turns out to be rather involved because it requires the general decision theoretic setup developed in Section 3.3 on base of L. Le Cam's work, important topological results from Section 4.2 and a strong result from the theory of vector lattices[6].

[5] cf. Section 3.2

[6] a Hahn-Banach-type theorem for M-Spaces; cf. Lemma 5.11

5.3.1.2 Main results and outline of the proof

Throughout this subsection, Θ is any index set, $\mathcal{X}$ is a set with algebras $\mathcal{A}$ and $\mathcal{A}'$ such that

$$\mathcal{A} \ \subset \ \mathcal{A}'$$

$\overline{\Pi}$ is a coherent upper prevision on $\mathcal{L}_\infty(\Theta, 2^\Theta)$ with credal set $\mathcal{P}$.
$(\overline{P}_\theta)_{\theta\in\Theta}$ is an imprecise model on $(\mathcal{X},\mathcal{A})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of credal sets. For every $\theta \in \Theta$, $\overline{P}_\theta'$ denotes the natural extension of $\overline{P}_\theta$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ and $\mathcal{M}_\theta'$ denotes its credal set on $(\mathcal{X},\mathcal{A}')$.
Finally, $(\mathbb{D},\mathcal{D})$ is a decision space and

$$W \ : \ \Theta \times \mathbb{D} \ \to \ \mathbb{R}\,, \qquad (\theta, t) \ \mapsto \ W_\theta(t)$$

is a loss function such that $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$.
Recall from Section 3.1 that the Bayes risk of a randomized decision function τ on $(\mathcal{X},\mathcal{A})$ is denoted by

$$R_{\overline{\Pi}}\big((\overline{P}_\theta)_{\theta\in\Theta}, \tau, W\big)$$

and the Bayes risk of a randomized decision function τ' on $(\mathcal{X},\mathcal{A}')$ is denoted by

$$R_{\overline{\Pi}}\big((\overline{P}_\theta')_{\theta\in\Theta}, \tau', W\big)$$

The following lemma exposes a simple but important fact:

Lemma 5.5 *Every randomized decision function τ on $(\mathcal{X},\mathcal{A})$ is also a randomized decision function on $(\mathcal{X},\mathcal{A}')$ and*

$$R_{\overline{\Pi}}\big((\overline{P}_\theta)_{\theta\in\Theta}, \tau, W\big) \ = \ R_{\overline{\Pi}}\big((\overline{P}_\theta')_{\theta\in\Theta}, \tau, W\big)$$

Proof: Let τ be a randomized decision function on $(\mathcal{X},\mathcal{A})$. Put

$$\tau_\bullet[h] \ : \ \mathcal{X} \ \to \ \mathbb{R}\,, \qquad x \ \mapsto \ \tau_x[h] \ = \ \int_{\mathbb{D}} h(t)\, \tau_x(dt)$$

for every $h \in \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$. Since

$$\tau_\bullet[I_D] \ \in \ \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \ \subset \ \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \qquad \forall\, D \in \mathcal{D}$$

τ is also a randomized decision function on $(\mathcal{X},\mathcal{A}')$ and

$$\begin{aligned}
R_{\overline{\Pi}}\big((\overline{P}_\theta')_{\theta\in\Theta}, \tau, W\big) \ &= \ \sup_{\pi\in\mathcal{P}} \int_\Theta \sup_{P_\theta'\in\mathcal{M}_\theta'} \int_{\mathcal{X}} \tau_x[W_\theta]\, P_\theta'(dx')\, \pi(d\theta) \ = \\
&= \ \sup_{\pi\in\mathcal{P}} \int_\Theta \overline{P}_\theta' \big[\tau_\bullet[W_\theta]\big] \ \pi(d\theta) \ = \ \sup_{\pi\in\mathcal{P}} \int_\Theta \overline{P}_\theta \big[\tau_\bullet[W_\theta]\big] \ \pi(d\theta) \ = \\
&= \ \sup_{\pi\in\mathcal{P}} \int_\Theta \sup_{P_\theta\in\mathcal{M}_\theta} \int_{\mathcal{X}} \tau_x[W_\theta]\, P_\theta(dx)\, \pi(d\theta) \ = \\
&= \ R_{\overline{\Pi}}\big((\overline{P}_\theta)_{\theta\in\Theta}, \tau, W\big)
\end{aligned}$$

□

According to Subsection 3.3.1, the randomized decision functions on $(\mathcal{X},\mathcal{A})$ correspond to the ordinary randomizations from $(\mathcal{X},\mathcal{A})$ to $(\mathbb{D},\mathcal{D})$. So, $\mathcal{T}_0(\mathcal{X},\mathbb{D})$ may also denote the set of all randomized decision functions on $(\mathcal{X},\mathcal{A})$. Accordingly, $\mathcal{T}_0'(\mathcal{X},\mathbb{D})$ may denote the (larger) set of all randomized decision functions on $(\mathcal{X},\mathcal{A}')$.

Now, we can state the main theorem of the present subsection:

Theorem 5.6 *Assume that Θ is a finite set. Then:*

a) [Γ-*minimax*]

$$\inf_{\tau'\in\mathcal{T}_0'(\mathcal{X},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{P}_\theta')_{\theta\in\Theta},\tau',W\big) \;=\; \inf_{\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{P}_\theta')_{\theta\in\Theta},\tau,W\big)$$

Furthermore, a randomized decision function on $(\mathcal{X},\mathcal{A})$ which is Γ-minimax over all randomized decision functions on $(\mathcal{X},\mathcal{A})$ is also Γ-minimax over all randomized decision functions on $(\mathcal{X},\mathcal{A}')$.

b) [*E-admissibility*] *Take any $\tilde{\tau}\in\mathcal{T}_0(\mathcal{X},\mathbb{D})$ so that there is some $\pi\in\mathcal{P}$ where*

$$R_\pi\big((\overline{P}_\theta)_{\theta\in\Theta},\tilde{\tau},W\big) \;=\; \inf_{\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big)$$

Then,

$$R_\pi\big((\overline{P}_\theta')_{\theta\in\Theta},\tilde{\tau},W\big) \;=\; \inf_{\tau'\in\mathcal{T}_0'(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}_\theta')_{\theta\in\Theta},\tau',W\big)$$

That is, a randomized decision function on $(\mathcal{X},\mathcal{A})$ which is E-admissible over all randomized decision functions on $(\mathcal{X},\mathcal{A})$ is also E-admissible over all randomized decision functions on $(\mathcal{X},\mathcal{A}')$.

As already stated above, the proof of Theorem 5.6 is rather involved. Nevertheless, it is based on a simple idea, which is presented in the following:

Let $T:\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')\to\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ be a map so that

- T is linear
- T is positive: $T(f')\geq 0 \quad \forall\, f'\geq 0$

and so that

- $T(f)=f \quad \forall\, f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$

Such a map does not always exist, but for a start, let us assume that such a map T would exist. Example 5.9 below presents a situation where T exists and can explicitly be specified. Furthermore, it contains an example where T cannot exist.

Now, let τ' be a randomized decision function

$$\tau' \;:\; \mathcal{X}\times\mathcal{D}\;\to\;\mathbb{R}\,, \qquad (x,D)\;\mapsto\;\tau_x'(D)$$

such that $\tau_\bullet'(D): x\mapsto\tau_x'(D)$ is an element of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. By use of T, it is easy to show, that there is a randomized decision function

$$\tau \;:\; \mathcal{X}\times\mathcal{D}\;\to\;\mathbb{R}\,, \qquad (x,D)\;\mapsto\;\tau_x(D)$$

such that $\tau_\bullet(D) : x \mapsto \tau_x(D)$ is an element of $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ and

$$R\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tau, W\big) \;\leq\; R\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tau', W\big) \tag{5.24}$$

To this end, note that T defines a finitely additive Markov kernel κ in the following way:

$$\kappa \;:\; \mathcal{X}\times\mathcal{A}' \;\to\; \mathbb{R}\,, \qquad (x, A') \;\mapsto\; \kappa_x(A') \;=\; T\big(I_{A'}\big)(x)$$

Next, put

$$\tau \;:\; \mathcal{X}\times\mathcal{D} \;\to\; \mathbb{R}\,, \qquad (x, D) \;\mapsto\; \tau_x(D) \;=\; \int_{\mathcal{X}} \tau'_{x'}(D)\; \kappa_x(dx')$$

It is easy to see that τ is a (finitely additive) Markov kernel $\tau : \mathcal{X}\times\mathcal{D} \to \mathbb{R}$ such that

$$\tau_x(D) \;=\; T\big(\tau'_\bullet(D)\big)(x) \qquad \forall\, x\in\mathcal{X}\,, \quad \forall\, D\in\mathcal{D}$$

and, therefore, $\tau_\bullet(D) : x \mapsto \tau_x(D)$ is an element of $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ for every $D\in\mathcal{D}$. Hence, τ is a randomized decision function on $(\mathcal{X}, \mathcal{A})$.
That is, T turns a randomized decision function τ' on $(\mathcal{X}, \mathcal{A}')$ into a randomized decision function τ on $(\mathcal{X}, \mathcal{A})$. It only remains to proof that the Bayes risk of τ is not larger than the Bayes risk of τ' – i.e. (5.24):
Note that

$$\rho(P)[f'] \;:=\; \int_{\mathcal{X}}\int_{\mathcal{X}} f'(x')\; \kappa_x(dx')\, P(dx) \;=\; P\big[T(f')\big]$$
$$\forall\, f'\in\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')\,, \quad \forall\, P\in \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A})$$

leads to a well defined map

$$\rho \;:\; \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}) \;\to\; \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}')$$

such that

$$P_\theta[f] \;\leq\; \overline{P}_\theta[f] \qquad \forall\, f\in\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$$

implies

$$\rho(P_\theta)[f'] \;\leq\; \overline{P}'_\theta[f'] \qquad \forall\, f'\in\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$$

In other words, we have

$$P_\theta\in\mathcal{M}_\theta \qquad \Rightarrow \qquad \rho(P_\theta)\in\mathcal{M}'_\theta \tag{5.25}$$

where $\mathcal{M}_\theta$ denotes the credal set of $\overline{P}_\theta$ and $\mathcal{M}'_\theta$ denotes the credal set of $\overline{P}'_\theta$.
Finally,

$$\begin{aligned}
\sup_{P'_\theta\in\mathcal{M}'_\theta} \int_{\mathcal{X}}\int_{\mathcal{D}} W_\theta(t)\,\tau_x(dt)\, P'_\theta \;&=\; \sup_{P_\theta\in\mathcal{M}_\theta} \int_{\mathcal{X}}\int_{\mathcal{D}} W_\theta(t)\,\tau_x(dt)\, P_\theta(dx) \;= \\
&= \sup_{P_\theta\in\mathcal{M}_\theta} \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{D}} W_\theta(t)\,\tau'_{x'}(dt)\,\kappa_x(dx')\, P_\theta(dx) \;= \\
&= \sup_{P_\theta\in\mathcal{M}_\theta} \int_{\mathcal{D}} W_\theta(t)\,\tau'_{x'}(dt)\;\big[\rho(P_\theta)\big](dx') \;\leq \\
&\overset{(5.25)}{\leq} \sup_{P'_\theta\in\mathcal{M}'_\theta} \int_{\mathcal{D}} W_\theta(t)\,\tau'_{x'}(dt)\; P'_\theta(dx')
\end{aligned}$$

implies (5.24). That is, we have already proven Theorem 5.6 a) – under the assumption that the map T exists. Unfortunately, T does not need to exist. Indeed, it is enough to consider $\mathcal{A} = \mathbb{B}$ as shown by Example 5.9 b). Example 5.9 a) presents a concrete example where T does exist.

In view of Subsection 3.3.1, T defines – if it exists – an ordinary randomization ρ. As we will see below, the key result in the proof of Theorem 5.6 is the fact that a suitable generalization of T – namely a *generalized* randomization – always exists. This is the content of Lemma 5.11 below, which is strongly based on the theory of vector lattices.

One may wonder if the undesirable assumption that Θ has to be a finite set in Theorem 5.6 is necessary. Essentially, this assumption is not necessary but it makes it possible to formulate Theorem 5.6 in terms of randomized decision functions (i.e. ordinary decision procedures) which certainly is more comfortable for most readers. The reader who is in complete accordance with L. Le Cam and thinks that generalized decision procedures are just as well (cf. also Section 3.4) may dispense with any assumption on Θ. This can be seen by the following Theorem 5.7. In fact, we will use Theorem 5.7 in order to proof Theorem 5.6.

Theorem 5.7 *Let Θ be any index set. Then:*
For every generalized decision procedure $\sigma' \in \mathcal{T}'(\mathcal{X},\mathbb{D})$, there is a generalized decision procedure $\sigma \in \mathcal{T}(\mathcal{X},\mathbb{D})$ such that the risk function of σ is not larger than the risk function of σ' in every $\theta \in \Theta$ – i.e.

$$\sup_{P_\theta \in \mathcal{M}_\theta} \sigma(P_\theta)[W_\theta] \;\leq\; \sup_{P'_\theta \in \mathcal{M}'_\theta} \sigma'(P'_\theta)[W_\theta] \qquad \forall\, \theta \in \Theta$$

In accordance with $\mathcal{T}_0(\mathcal{X},\mathbb{D})$ and $\mathcal{T}_0'(\mathcal{X},\mathbb{D})$ from above, $\mathcal{T}(\mathcal{X},\mathbb{D})$ denotes the set of all generalized decision procedures in case of the sample space $(\mathcal{X},\mathcal{A})$ and $\mathcal{T}'(\mathcal{X},\mathbb{D})$ denotes the set of all generalized decision procedures in case of the sample space $(\mathcal{X},\mathcal{A}')$. Risk functions for generalized decision procedures have already been defined in Subsection 3.3.1.2.

Another possibility to get rid of the assumption that Θ has to be finite is to assume that $\mathcal{A}$ is finite. This is because the map T always exists if $\mathcal{A}$ is finite. Especially, this case is relevant for discretizations; cf. Section 5.4.

Theorem 5.8 *Let Θ be any index set and let $\mathcal{A}$ be finite. Then:*
For every randomized decision function $\tau' \in \mathcal{T}_0'(\mathcal{X},\mathbb{D})$, there is a randomized decision function $\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})$ such that the risk function of τ is not larger than the risk function of τ' in every $\theta \in \Theta$ – i.e.

$$\sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau_x(dt)\, P'_\theta(dx) \;\leq\; \sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau'_x(dt)\, P'_\theta(dx)$$

for every $\theta \in \Theta$.

The present subsection closes with the repeatedly mentioned example concerning existence of the map T. The following subsection is concerned with the proofs of the above theorems.

Example 5.9

a) Take $\mathcal{X} = \mathbb{R}$, $\mathcal{A}' = \mathbb{B}$ and let $\mathcal{A}$ be the σ-algebra which is generated by the sets $[k, k+1)$, $k \in \mathbb{Z}$. Then,

$$T: \; f' \;\mapsto\; \sum_{k \in \mathbb{Z}} f'(k) I_{[k,k+1)}$$

is a linear, positive map $T : \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}') \to \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ such that $T(f) = f$ for every $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}) \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$.

b) Take $\mathcal{X} = \mathbb{R}$, $\mathcal{A} = \mathbb{B}$ and let $\mathcal{A}'$ be the power set of $\mathbb{R}$. Then, such a map T as in a) cannot exist.

In order to see this, note that $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ and $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ are M-spaces. Furthermore, $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ is Dedekind complete. Now, assume that T would exist and let $F \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ be a majorized subset of $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ – i.e. $\exists f_0 \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}) : \quad f \leq f_0 \quad \forall f \in F$. Hence, F is also a majorized subset of $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ and Dedekind completeness of $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ implies the existence of a supremum $h' \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ of F. Next, put $h = T(h')$ and it follows from positivity of T that h is a supremum of F in $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$.
As a consequence, existence of T would imply that $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ was order complete, too. However, $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ is not *order complete.*[7]

5.3.1.3 Proof of the main result

Let $\mathcal{X}$ again be a set with algebras $\mathcal{A}$ and $\mathcal{A}'$ such that $\mathcal{A} \subset \mathcal{A}'$. As presented in the previous subsection, the proof of Theorem 5.6 would be rather clear if a certain map

$$T \;:\; \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}') \;\to\; \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$$

would always exist. Though this is not possible, Lemma 5.11 states that a suitable generalization of T does always exist. Since T corresponds to an ordinary randomization, it is not surprising that the generalization is given by a generalized randomization. This generalization is called *extending transition* in Definition 5.10.

Definition 5.10 *A map*

$$\rho \;:\; \mathrm{ba}(\mathcal{X}, \mathcal{A}) \;\to\; \mathrm{ba}(\mathcal{X}, \mathcal{A}')$$

is called extending transition from $\mathcal{A}$ to $\mathcal{A}'$ *if*

- *ρ is linear*
- *ρ is positive: $\rho(\mu) \geq 0 \quad \forall \mu \geq 0$*

and

$$\rho(\mu)[f] \;=\; \mu[f] \qquad \forall f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}), \quad \forall \mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A}) \tag{5.26}$$

[7] In order to see this, take any $B' \subset \mathbb{R}$ such that $B' \notin \mathbb{B} = \mathcal{A}$ and put

$$F = \big\{ I_{\{x\}} \;\big|\; x \in B \big\}$$

F is majorized by $f_0 \equiv 1$ Assume that h would be a supremum of F in $\mathcal{L}_\infty(\mathcal{X}, \mathbb{B})$. Then, it follows that h is an indicator function – i.e., there is a set $B \in \mathbb{B}$ such that $h = I_B$. Next, $B' \notin \mathbb{B}$ implies $B' \subsetneq B$. Take any $x \in B \setminus B'$ and put $\hat{h} = I_{B \setminus \{x\}} \in \mathcal{L}_\infty(\mathcal{X}, \mathbb{B})$. Then, $\hat{h}$ is a majorant of F such that $\hat{h} \lneq h$. Hence, h is *not* a supremum.

It is easy to see that an extending transition is in fact a transition according to Definition 3.34:
Assertion (5.26) implies that

$$\rho(\mu)[I_{\mathcal{X}}] \;=\; \mu[I_{\mathcal{X}}] \qquad \forall\,\mu \in \mathrm{ba}(\mathcal{X},\mathcal{A})$$

and, therefore, ρ is a generalized randomization from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{X},\mathcal{A}')$. According to Proposition 3.36, the generalized randomizations from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{X},\mathcal{A}')$ are precisely the transitions from $\mathrm{ba}(\mathcal{X},\mathcal{A})$ to $\mathrm{ba}(\mathcal{X},\mathcal{A}')$.
Furthermore, (5.26) says that $\rho(\mu)$ is always an extension of μ from $\mathcal{A}$ to $\mathcal{A}'$. This justifies the term "extending transition".

In contrast to T, an extending transition ρ does always exist:

Lemma 5.11 *There is an extending transition*

$$\rho:\; \mathrm{ba}(\mathcal{X},\mathcal{A}) \;\longrightarrow\; \mathrm{ba}(\mathcal{X},\mathcal{A}')$$

The proof of Lemma 5.11 is strongly based on the theory of vector lattices. For the theory of vector lattices, confer e.g. Schaefer (1974) and Section 8.1 in the Appendix.

Proof:

[1] Firstly, it is shown that there is a positive linear operator $S \,:\, \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \to \big(\mathrm{ba}(\mathcal{X},\mathcal{A})\big)^*$ so that

$$S(f) = \Lambda_f \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})\,, \qquad \|S\| \;=\; 1$$

where $\Lambda_f[\mu] = \mu[f]\;\;\forall\,\mu \in \mathrm{ba}(\mathcal{X},\mathcal{A})$ and $\big(\mathrm{ba}(\mathcal{X},\mathcal{A})\big)^*$ denotes the dual space of $\mathrm{ba}(\mathcal{X},\mathcal{A})$:

$G_0 := \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ is a Banach sublattice of the Banach lattice $G := \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. According to (Schaefer, 1974, p. 114), $\mathrm{ba}(\mathcal{X},\mathcal{A})$ is an abstract L-space. So, it follows from Schaefer (1974, Prop. 5.5 and Prop. 9.1) that $E := \big(\mathrm{ba}(\mathcal{X},\mathcal{A})\big)^*$ is an order complete M space with unit.
Note that $S_0 \,:\, G_0 \;\to\; E\,,\;\; f \;\mapsto\; \Lambda_f$ is a positive linear operator where $\|S_0\| \;=\; 1$. According to Schaefer (1974, Cor. 7.10.3), S_0 can be extended to a positive linear operator S on G such that $\|S\| = \|S_0\| = 1$. Hence, [1].

[2] Next, it is shown that $\rho:\; \mathrm{ba}(\mathcal{X},\mathcal{A}) \to \mathrm{ba}(\mathcal{X},\mathcal{A}'),\;\; \mu \mapsto \rho(\mu)$ where

$$\rho(\mu)[f'] \;=\; S(f')[\mu] \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

is an extending transition:

The properties of S imply that ρ is a linear, positive operator. Furthermore,

$$\rho(\mu)[f] = S(f)[\mu] = S_0(f)[\mu] = \mu[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}),\;\; \forall\,\mu \in \mathrm{ba}(\mathcal{X},\mathcal{A})$$

□

Now, it is possible to proof Theorem 5.7.

Proof of Theorem 5.7: For every generalized randomization $\sigma' \in \mathcal{T}'(\mathcal{X},\mathbb{D})$, put $\sigma := \sigma' \circ \rho$ where ρ is the extending transition from $\mathcal{A}$ to $\mathcal{A}'$ according to Lemma 5.11. Then, $\sigma \in \mathcal{T}(\mathcal{X},\mathbb{D})$ because ρ is a generalized randomization from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{X},\mathcal{A}')$. Fix any $\theta \in \Theta$. Since $\overline{P}'_\theta$ is the natural extension of $\overline{P}_\theta$, the credal set of $\overline{P}'_\theta$ is given by

$$\mathcal{M}'_\theta \;=\; \left\{ P'_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \;\middle|\; P'_\theta[f] \le \overline{P}_\theta[f] \quad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}) \right\}$$

according to Proposition 2.13. Therefore, assertion

$$P_\theta \;\in\; \mathcal{M}_\theta \qquad \Rightarrow \qquad \rho(P_\theta) \;\in\; \mathcal{M}'_\theta \tag{5.27}$$

follows from

$$\rho(P_\theta)[f] \;=\; P_\theta[f] \;\le\; \overline{P}_\theta[f] \qquad \forall\, f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

Finally,

$$\begin{aligned} \sup_{P_\theta \in \mathcal{M}_\theta} \sigma(P_\theta)[W_\theta] \quad &= \quad \sup_{P_\theta \in \mathcal{M}_\theta} \sigma' \circ \rho(P_\theta)[W_\theta] = \sup_{P_\theta \in \mathcal{M}_\theta} \sigma'\big(\rho(P_\theta)\big)[W_\theta] \;\le \\ &\overset{(5.27)}{\le} \quad \sup_{P'_\theta \in \mathcal{M}'_\theta} \sigma'(P_\theta)[W_\theta] \end{aligned}$$

□

Next, Theorem 5.6 is proven by use of Theorem 5.7:

Proof of Theorem 5.6: For every $\pi \in \mathcal{P}$,

$$R_\pi\big((\overline{P}'_\theta)_{\theta\in\Theta}, \sigma', W\big) \;=\; \sum_{\theta\in\Theta} \pi(\{\theta\}) \sup_{P'_\theta \in \mathcal{M}'_\theta} \sigma'(P'_\theta)[W_\theta]$$

is the Bayes risk of a decision procedure $\sigma' \in \mathcal{T}'(\mathcal{X},\mathbb{D})$ with respect to the (precise) prior π . It follows from Theorem 5.7 and Lemma 5.5 that

$$\inf_{\sigma' \in \mathcal{T}'(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}'_\theta)_{\theta\in\Theta}, \sigma', W\big) \;=\; \inf_{\sigma \in \mathcal{T}(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}'_\theta)_{\theta\in\Theta}, \sigma, W\big)$$

Hence, Proposition 4.5 implies

$$\inf_{\tau' \in \mathcal{T}'_0(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tau', W\big) \;=\; \inf_{\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tau, W\big) \tag{5.28}$$

for every $\pi \in \mathcal{P}$.

a) Put

$$\Gamma(\pi,\tau) \;:=\; -R_\pi\big((\overline{P}_\theta)_{\theta\in\Theta}, \tau, W\big) \;=\; -\sum_{\theta\in\Theta} \pi(\{\theta\}) \sup_{P_\theta \in \mathcal{M}_\theta} \tau(P_\theta)[W_\theta]$$

for every $\pi \in \mathcal{P}$ and every $\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})$.

Hence, $\tau \mapsto \Gamma(\pi,\tau)$ is concave for every $\pi \in \mathcal{P}$ and $\mathcal{P}$ is $\mathcal{L}_\infty(\Theta,2^\Theta)$ - compact according to Corollary 2.16. Furthermore, $\pi \mapsto \Gamma(\pi,\tau)$ is convex and $\mathcal{L}_\infty(\Theta,2^\Theta)$ - continuous for every $\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})$. Hence

$$\sup_{\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})} \inf_{\pi \in \mathcal{P}} \Gamma(\pi,\tau) \;=\; \inf_{\pi \in \mathcal{P}} \sup_{\tau \in \mathcal{T}_0(\mathcal{X},\mathbb{D})} \Gamma(\pi,\tau)$$

according to Fan (1953, Theorem 1). That is,

$$\inf_{\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})}\ \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}_\theta)_{\theta\in\Theta},\tau,W\big)$$

An analogous proof for $\mathcal{A}'$ shows

$$\inf_{\tau'\in\mathcal{T}_0'(\mathcal{X},\mathbb{D})}\ \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{P}_\theta')_{\theta\in\Theta},\tau',W\big) \;=\; \sup_{\pi\in\mathcal{P}}\ \inf_{\tau'\in\mathcal{T}_0'(\mathcal{X},\mathbb{D})} R_\pi\big((\overline{P}_\theta')_{\theta\in\Theta},\tau',W\big)$$

Finally, Part a) follows from the latter equations, (5.28) and Lemma 5.5.

b) This is a direct consequence of (5.28).

□

Finally, it only remains to proof Theorem 5.8.

Proof of Theorem 5.8: According to Definition 3.6, the randomized decision functions correspond to ordinary randomizations. Therefore, the proof may be formulated in terms of ordinary randomizations.

Firstly, we have to show that a map T as discussed in Subsection 5.3.1.2 exists.

Since $\mathcal{A}$ is finite, there is a partition $\{A_1,\dots,A_m\}\subset\mathcal{A}$ of $\mathcal{X}$ so that $A_j\neq\emptyset\ \ \forall j=1,\dots,m$ and so that every $A\in\mathcal{A}$ is a union of some elements of $\{A_1,\dots,A_m\}$. For every $j=1,\dots,n$, choose any $x_j\in A_j$. Next, put

$$T\ :\ \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')\ \to\ \mathcal{L}_\infty(\mathcal{X},\mathcal{A}),\qquad f'\ \mapsto\ T(f')\ =\ \sum_{j=1}^{m} f'(x_j)\cdot I_{A_j}$$

Obviously, T is linear and positive. In addition,

$$T(f)\ =\ f\qquad \forall\, f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

is fulfilled because every function $f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ is of form

$$f\ =\ \sum_{i=1}^{m}\alpha_j I_{A_j}\,,\qquad \alpha_1,\dots,\alpha_m\ \in\ \mathbb{R}$$

According to Proposition 3.11, T defines an ordinary randomization

$$\rho\ :\ \operatorname{ba}(\mathcal{X},\mathcal{A})\ \to\ \operatorname{ba}(\mathcal{X},\mathcal{A}')\,,\qquad \mu\ \mapsto\ \rho(\mu)$$

via $\rho(\mu)[f']=\mu\big[T(f')\big]$ for every $f'\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ and $\mu\in\operatorname{ba}(\mathcal{X},\mathcal{A})$.

Obviously, ρ is an extending transition. Therefore, the remaining part of the proof is very similar to the proof of Theorem 5.7:

For every ordinary randomization $\tau'\in\mathcal{T}_0'(\mathcal{X},\mathbb{D})$, put $\tau:=\tau'\circ\rho$. Then, it follows from Proposition 3.11 c) that $\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})$. Fix any $\theta\in\Theta$. Since $\overline{P}_\theta'$ is the natural extension of $\overline{P}_\theta$, assertion

$$P_\theta\ \in\ \mathcal{M}_\theta\qquad\Rightarrow\qquad \rho(P_\theta)\ \in\ \mathcal{M}_\theta' \tag{5.29}$$

follows from

$$\rho(P_\theta)[f]\ =\ P_\theta[f]\ \le\ \overline{P}_\theta[f]\qquad \forall\, f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$$

Finally,

$$\begin{aligned}\sup_{P_\theta\in\mathcal{M}_\theta}\sigma(P_\theta)[W_\theta] &= \sup_{P_\theta\in\mathcal{M}_\theta}\sigma'\circ\rho(P_\theta)[W_\theta] = \sup_{P_\theta\in\mathcal{M}_\theta}\sigma'\big(\rho(P_\theta)\big)[W_\theta] \le \\ &\overset{(5.29)}{\le} \sup_{P'_\theta\in\mathcal{M}'_\theta}\sigma'(P'_\theta)[W_\theta]\end{aligned}$$

□

5.3.2 Reduction of the sample space

The results given in Subsection 5.3.1.2 shows that it is not necessary to extend a decision problem to a larger sample space in order to check optimality of a decision procedure. In addition, the results offer the opportunity to reduce the sample space in some cases. In this way, the results may be used to simplify some decision problems drastically.

In the following, two examples are discussed where this is possible:

Let Θ be an index set, let $\mathcal{X}$ be a set with algebra $\mathcal{A}'$ and let $(\overline{P}'_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{X},\mathcal{A}')$.
Let $\overline{\Pi}$ be a coherent upper prevision on $\mathcal{L}_\infty(\Theta,2^\Theta)$, let $(\mathbb{D},\mathcal{D})$ be a decision space and

$$W\ :\ \Theta\times\mathbb{D}\ \longrightarrow\ \mathbb{R}\,,\qquad (\theta,t)\ \mapsto\ W_\theta(t)$$

a loss function.
1. Assume that Θ is finite and that there is a set $\mathcal{K}\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ such that the credal set of $\overline{P}'_\theta$ is given by

$$\mathcal{M}'_\theta\ =\ \big\{P_\theta\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\ \big|\ P'_\theta[f]\le\overline{P}_\theta[f]\quad\forall\, f\in\mathcal{K}\big\}$$

The task is: Find an optimal randomized decision function $\tilde{\tau}'$ on $(\mathcal{X},\mathcal{A}')$.
So far, this is a standard situation in decision theory under imprecise probabilities. Next, let $\mathcal{A}$ be the smallest algebra on $\mathcal{X}$ such that

$$f\ \in\ \mathcal{L}_\infty(\mathcal{X},\mathcal{A})\qquad\forall\, f\in\mathcal{K}$$

If $\mathcal{K}$ is not too large, $\mathcal{A}$ may be considerably smaller than $\mathcal{A}'$. Then, it follows from Theorem 5.6 that it is enough to consider randomized decision functions on the smaller sample space $(\mathcal{X},\mathcal{A})$.
For example, assume that $\mathcal{K}$ is a finite set of simple functions on $(\mathcal{X},\mathcal{A}')$. Then, $\mathcal{A}$ is always a finite algebra. Therefore, the infinite sample space $(\mathcal{X},\mathcal{A}')$ may be reduced to a finite sample space[8] $(\mathcal{X},\mathcal{A})$ in this case – and the decision problem is accessible for methods concerning finite spaces such as linear programming! Especially, this example applies for discretizations; confer Section 5.4.

2. Now, assume that the credal set of $\overline{P}'_\theta$ is given by

$$\mathcal{M}'_\theta\ =\ \big\{P_\theta\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\ \big|\ P_\theta(C_j)\le\overline{P}_\theta(C_j)\quad\forall\, j\in\{1,\dots,m\}\,\big\}$$

[8] Finiteness of $\mathcal{A}$ implies that the sample space $(\mathcal{X},\mathcal{A})$ may be considered as a finite one because, in this case, $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})\cong\mathbb{R}^n$ for some suitable $n\in\mathbb{N}$.

where $C_1, \ldots, C_m \in \mathcal{A}'$, $m \in \mathbb{N}$. Of course, this assumption is very restrictive. However, this is an important model which is frequently used in the theory of imprecise probabilities. Then, let $\mathcal{A}$ be the smallest (σ-)algebra which contains $C_1, \ldots, C_m$ and it follows that $\mathcal{A}$ is finite. Therefore, Theorem 5.6 applies and we may – without loosing anything – work with the finite sample space $(\mathcal{X}, \mathcal{A})$ instead of the (possibly) infinite sample space $(\mathcal{X}, \mathcal{A}')$.

5.4 Application: Discretizations

5.4.1 The meaning of discretizing in (statistical) decision theory

Though the theoretical evaluations of Section 5.2 and Section 5.3 both are independently interesting on its own, together they lead to important tools in applications. As already mentioned in the introductory Section 5.1, the results of Section 5.2 and Section 5.3 are the cornerstones of discretization in applications of (data-based) decision theory.

Discretization is a crucial topic in the theory of imprecise probabilities which has recently been considered in Troffaes (2008) and Obermeier and Augustin (2007). The fundamental importance of discretization in the theory of imprecise probabilities is explained in (Obermeier and Augustin, 2007, p. 327):

> "Classical statistical models typically are based on parametric, absolutely continuous probability distributions on the real line. Handling extensions of these models in the imprecise probability framework, quite often becomes very demanding from the computational point of view, and then approximative techniques are the best one can hope for, the more as also in classical statistics many integrals of less smooth functions can be only obtained numerically. A natural idea in this context is discretization, in order to make available powerful algorithms (...) that explicitly rely on finite spaces to obtain approximate solutions in this generalized setting. However, such discretizations need some care; for more than hundred years, since the work of Sheppard (...) [Sheppard (1898)] at the end of the nineteenth century, statisticians have been well aware that analysis based on rounded data may be severely biased, and so discretization by mere rounding or other ad-hoc techniques is a bad advice."

While Troffaes (2008) and Obermeier and Augustin (2007) consider decision theory which is not explicitly data-based, the following evaluation is probably the first one which deals with discretizations in decision theory (under imprecise probabilities) which is explicitly data-based. Accordingly, Troffaes (2008) and Obermeier and Augustin (2007) consider discretizations of Θ whereas the following evaluation is mainly concerned with discretizations of the sample space $(\mathcal{X}, \mathcal{A}')$. Discretizing Θ is a fundamental topic in general decision theory. However, if we focus on applications in statistics (i.e. in the special case of statistical decision theory), discretizations of the sample space seem to be even more important than discretizations of Θ:

Most part of statistics is about testing and estimating. In case of testing, discretizing Θ is not necessary (because $\Theta = \{0; 1\}$ already is discrete) but discretizing the sample space is a crucial issue.
In case of estimating, the set of all decisions $\mathbb{D}$ is equal to Θ and choosing decision $t = \hat{\theta}$

means that our estimation for the true parameter θ is $t = \hat{\theta}$. Now, what does discretizing Θ mean in this case? For example, take $\Theta = [0, 1]$ and consider the equidistant discretization

$$\mathcal{H} \;=\; \Big\{ \big[0, \tfrac{1}{10}\big]; \big(\tfrac{1}{10}, \tfrac{2}{10}\big]; \; \big(\tfrac{2}{10}; \tfrac{3}{10}\big]; \; \ldots \; ; \; \big(\tfrac{9}{10}; 1\big] \Big\}$$

Such a discretization changes the set of parameters and precisely means that we do not want to estimate the true θ any more but we want to estimate the interval

$$H \;=\; \big(\tfrac{k-1}{10}\,; \tfrac{k}{10}\,\big]$$

which contains the true parameter θ. Accordingly, the set of all decisions is equal to $\mathcal{H}$ now.

In order to explain the meaning of discretizations of Θ, (Troffaes, 2008, § 1) states:

> "(...) we must resort to computers, and these cannot handle gambles on infinite spaces, let alone arbitrary infinite sets of probabilities. Hence, in that case we must approximate our infinite sets by finite ones. By taking the finite sets sufficiently large, hopefully the approximation reflects the true result accurately."

This is true for the setup used in Troffaes (2008) but it does not apply for estimating. In Troffaes (2008), Θ is discretized but the set of all decisions is not changed. By discretizing, Troffaes (2008) wants to approximately solve the original decision problem. In contrast, if Θ is discretized in a statistical estimation problem, not only the parameter space gets coarser but also the decision space (since the decision space is equal to the parameter space). In this way, discretizing Θ implies that also our decision theoretic purpose gets "coarser": Now, we do not want to estimate the true θ but the true "discretized θ". This is something different. We do not want to approximately solve the original decision problem now but we do want to exactly solve a coarser decision problem. So, in a sense, discretizing Θ may be considered rather as part of modeling than as part of *solving* an estimation problem.

By discretizing Θ, an estimating problem gets easier in two aspects:

- Discretizing enables the use of "finite methods" such as linear programming so that the problem gets tractable by computers.[9]

- As a side effect, the estimating problem gets easier in the sense that coarser parameters such as intervals $H = \big(\frac{k-1}{10}\,; \frac{k}{10}\,\big]$ can be estimated more efficiently than precise parameters such as $\theta \in \Theta = [0, 1]$.
  For example, consider the most extreme discretization of Θ, namely $\mathcal{H} = \{\Theta\}$. Undoubtfully, the only remaining parameter $H = \Theta$ may be estimated in an exceedingly efficient way.

Usually, Θ is a subset of $\mathbb{R}$ or $\mathbb{R}^k$ so that discretizing Θ will often lead to a parameter set which consists of intervals (or hyperrectangle in $\mathbb{R}^k$). This fits very well into the theory of imprecise probabilities because such interval valued parameters naturally avoid over-precise estimations. Even statisticians which do not agree with imprecise probabilities are

[9] Of course, this does not mean that computers are always able to solve the problems within human time scales. It only means that, in principle, computers can solve such problems.

aware of the fact that estimating should not be done in an over-precise way. This becomes apparent because statisticians usually refuse to record the outcome of an estimation with the highest accuracy computers can provide. Instead, the outcomes of estimations are "reasonably" rounded and this essentially means that only interval valued estimations are commonly accepted.

Since the focus of the present book lies on applications in statistics and, there, discretizing Θ is often not necessary (testing) or strongly interrelated to modeling (estimating), discretizations of Θ are hardly considered in this book. Instead, it is often assumed in the following that Θ already is a finite set. Of course, this is a restrictive assumption but the above reasoning may justify it at least to some extend. Subsection 6.2.2 only provides an ad hoc method for discretizing Θ in a certain estimation problem if Θ happens to be not discrete and it would be desirable to develop more sophisticated methods in the setup of Subsection 6.2.2.

As mentioned above, discretizing the sample space $(\mathcal{X},\mathcal{A}')$ is a crucial issue in statistical applications. Here, we require indeed that, by "taking the finite sets sufficiently large, hopefully the approximation reflects the true result accurately."[10] And, as already stated in (Obermeier and Augustin, 2007, p. 327), such discretizations have to be done carefully since simple rounding of the data may lead to bad results. It is the purpose of the following subsection to derive a well justified *and* practicable method for disrectizing $(\mathcal{X},\mathcal{A}')$ which is based on the theoretical evaluations of the preceeding sections.

5.4.2 A method for discretizing sample spaces

Setup and assumptions:

It is assumed that, for every $\theta\in\Theta$, there is a finite subset $\mathcal{K}_\theta\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ such that the credal set of $\overline{P}'_\theta$ is given by

$$\mathcal{M}'_\theta \;=\; \left\{P'_\theta\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\;\middle|\;P'_\theta[f]\le\overline{P}'_\theta[f]\quad\forall\, f\in\mathcal{K}_\theta\right\} \tag{5.30}$$

Furthermore, it is assumed that

$$\mathcal{K}\;:=\;\bigcup_{\theta\in\Theta}\mathcal{K}_\theta\qquad\text{is a finite set} \tag{5.31}$$

Finally, it is assumed, that, for every fixed $f\in\mathcal{K}$, there is a $d_f>0$ such that

$$\overline{P}'_\theta[f]-\underline{P}'_\theta[f]\quad\ge\quad d_f\qquad\quad\text{for every }\theta\in\Theta\text{ where }\mathcal{K}_\theta\ni f \tag{5.32}$$

Assumption (5.30) is crucial and rather restrictive – nevertheless, such imprecise models are quite important for practical applications as explained below.
The index set Θ is not assumed to be finite here. Instead, the considerably weaker Assumption (5.31) is sufficient. Of course, (5.33) is fulfilled if Θ is finite but it is also fulfilled if $\mathcal{K}_\theta$ does not depend on $\theta\in\Theta$.
If Θ is finite, then Assumption (5.32) coincides with the assumption

$$\overline{P}'_\theta[f]\;\ne\;\underline{P}'_\theta[f]\qquad\forall\, f\in\mathcal{K}_\theta \tag{5.33}$$

[10] (Troffaes, 2008, § 1)

In addition to Assumption (5.30), Assumption (5.33) is not restrictive at all because Section 5.2 tell us: Using models of form (5.30) which violate (5.33) is dangerous because these models are potentially most instable. Therefore, those models which violate (5.33) generally should be avoided anyway. If modeling led to a coherent upper prevision which violates (5.33), it should be replaced by a more cautious and stable coherent upper prevision according to (5.5).

For practical applications, it is important that the validity of these assumptions can easily be checked:

Usually, the validity of (5.30) and (5.31) directly results from modeling: A practitioner specifies concrete upper previsions for a finite number of functions $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ in order to get coherent upper previsions

$$\overline{P}'_\theta \ : \ \mathcal{K}_\theta \ \to \ \mathbb{R}\,, \qquad\qquad \theta \in \Theta \tag{5.34}$$

Next, these coherent upper previsions are extended on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ by the method of natural extension and this leads to an imprecise model which fulfills (5.30) and (5.31).
If Θ is infinite, then the infinite number of upper previsions in (5.34) has been specified by analytical arguments and, therefore, the validity of (5.32) has to be checked also by analytical arguments.
If Θ is finite, then the following proceeding can be applied:

For every $f_0 \in \mathcal{K}$ take a partition $\{B_1,\ldots,B_k\}$ of $\mathcal{X}$ and put

$$\overline{f} \ := \ \sum_{j=1}^{k} \sup_{x_j \in B_j} f(x_j)\cdot I_{B_j} \qquad \forall\, f \in \mathcal{K}$$

For every $\theta \in \Theta$ such that $\mathcal{K}_\theta \ni f_0$, solve the following linear programm:

$$\big(-\overline{f}_0(b_1)\,,\ \ldots\ ,-\overline{f}_0(b_k)\big)\cdot p \ \ \to \ \ \max \tag{5.35}$$

where

$$\big(\overline{f}(b_1)\,,\ \ldots\ ,\overline{f}(b_k)\big)\cdot p \ \ \le \ \ \overline{P}'_\theta[f] \qquad \forall\, f \in \mathcal{K}_\theta$$

and

$$p \in \mathbb{R}^k\,, \qquad p_j \ge 0 \ \ \forall\, j \in \{1,\ldots,k\}\,, \qquad p_1 + \ \ldots\ + p_k \ = \ 1$$

If the optimal value $l_{f_0,\theta}$ is not larger than $-\overline{P}'_\theta[f]$, start again with a finer partition.
If the optimal values $l_{f_0,\theta}$ are larger than $-\overline{P}'_\theta[f_0]$ for every $\theta \in \Theta$, put

$$d_{f_0} \ \ := \ \ \min\big\{\overline{P}'_\theta[f_0] + l_{f_0,\theta} \ \big|\ \theta \in \Theta \ : \ \mathcal{K}_\theta \ni f_0\big\}$$

Assumption (5.30) is fulfilled, if this procedure ends up with positive numbers d_f, $f \in \mathcal{K}$. This is a consequence of the following proposition which states that $\overline{P}'_\theta[f] + l_{f,\theta}$ is a lower bound on $\overline{P}'_\theta[f] - \underline{P}'_\theta[f]$. Of course, the finer partition $\{B_1,\ldots,B_l\}$ is, the better lower bound $\overline{P}'_\theta[f] + l_{f,\theta}$ usually is.

Proposition 5.12 *For a fixed $f_0 \in \mathcal{K}$ and a fixed $\theta \in \Theta$, let $l_{f_0,\theta}$ be the optimal value in the linear program (5.35). Then,*

$$\overline{P}'_\theta[f_0] + l_{f_0,\theta} \ \ \le \ \ \overline{P}'_\theta[f_0] - \underline{P}'_\theta[f_0]$$

Proof: Put

$$\hat{\mathcal{M}}'_\theta \;:=\; \big\{P'_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \;\big|\; P'_\theta\,\big[\,\overline{f}\,\big] \le \overline{P}'_\theta[f] \quad \forall\, f\in\mathcal{K}_\theta\big\}$$

The construction implies that the optimal value $l_{f_0,\theta}$ in the linear program (5.35) is equal to

$$l_{f_0,\theta} \;=\; \sup_{P'_\theta\in\hat{\mathcal{M}}'_\theta} P'_\theta\big[-\overline{f}_0\big]$$

Note that $f \;\le\; \overline{f}$ for every $f\in\mathcal{K}_\theta$. Hence, $\hat{\mathcal{M}}'_\theta \subset \mathcal{M}'_\theta$ and, therefore,

$$l_{f_0,\theta} \;=\; \sup_{P'_\theta\in\hat{\mathcal{M}}'_\theta} P'_\theta\big[-\overline{f}_0\big] \;=\; -\inf_{P'_\theta\in\hat{\mathcal{M}}'_\theta} P'_\theta\big[\overline{f}_0\big] \;\le\; -\underline{P}'_\theta\big[\overline{f}_0\big] \;\le\; -\underline{P}'_\theta[f_0]$$

□

Proceeding of the discretization

Recall the notation from the previous subsection and assume that $(\overline{P}'_\theta)_{\theta\in\Theta}$ is an imprecise model on the sample space $(\mathcal{X},\mathcal{A}')$ such that (5.30), (5.31) and (5.32) are fulfilled. Let $f_1,\dots,f_n$ be elements of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ such that

$$\{f_1,\,\dots\,,f_n\} \;=\; \mathcal{K} \;=\; \bigcup_{\theta\in\Theta}\mathcal{K}_\theta$$

and put $\mathcal{I}_\theta \;:=\; \big\{i\in\{1,\dots,n\}\;\big|\; f_i\in\mathcal{K}_\theta\big\} \quad \forall\,\theta\in\Theta\,.$

Proceed in the following way for any fixed $\varepsilon\in(0,1)$.

<u>*STEP 1:*</u> For every $i\in\{1,\dots,n\}$, take $d_i = d_{f_i}$ from (5.32) and put

$$\varepsilon_i \;:=\; \frac{\sup f_i - \inf f_i}{c\cdot d_i}\cdot\varepsilon \qquad \text{where} \quad c \;:=\; \sup_{\theta\in\Theta}\sum_{j\in\mathcal{I}_\theta}\frac{\sup f_j - \inf f_j}{d_j} \tag{5.36}$$

Note that the validity of

$$\sum_{j\in\{1,\dots,n\}}\frac{\sup f_j - \inf f_j}{d_j} \;\ge\; c \;\ge\; \frac{\sup f_i - \inf f_i}{d_i}$$

ensures $0<\varepsilon_i\le\varepsilon<1$. There is a $M\in\mathbb{N}$ such that

$$M-1 \;\le\; \frac{c}{\varepsilon} \;\le\; M \tag{5.37}$$

For every $i\in\{1,\dots,n\}$, put

$$b_i^{(j)} \;\;:=\;\; \inf f_i + \tfrac{j}{M}(\sup f_i - \inf f_i) \qquad \forall\, j\in\big\{0,1,2,\dots,M\big\}$$

and $A_i^{(1)} = f_i^{-1}\Big(\big[b_i^{(0)},\,b_i^{(1)}\big]\Big)$ and

$$A_i^{(j)} \;\;:=\;\; f_i^{-1}\Big(\big(b_i^{(j-1)},\,b_i^{(j)}\big]\Big) \;\;\in\;\; \mathcal{A}' \qquad \forall\, j\in\big\{2,\dots,M\big\} \tag{5.38}$$

Then, put

$$s_i \;:=\; \sum_{j=0}^{M} b_i^{(j)} I_{A_i^{(j)}} \qquad \text{and note that} \quad f_i \;\le\; s_i \;\le\; f_i + \varepsilon_i d_i \tag{5.39}$$

Let $\mathcal{A}$ be the smallest σ-algebra which contains $A_i^{(j)}$ for every $j \in \{1, \dots, M\}$ and every $i \in \{1, \dots, n\}$. Note that there is a finite partition

$$A_1\,,\;\dots\;,\;A_r \tag{5.40}$$

of $\mathcal{X}$ such that every element of $\mathcal{A}$ is the union of some elements of the partition $A_1, \dots, A_r$.

<u>*STEP 2:*</u> For every $\theta \in \Theta$, let $\overline{Q}_\theta$ be the coherent upper prevision on $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ which corresponds to the credal set

$$\mathcal{N}_\theta \;=\; \big\{\, Q_\theta \in \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}) \;\big|\; Q_\theta[s_i] \;\le\; \overline{P}'_\theta[f_i] + \varepsilon_i d_i \quad \forall\, i \in \mathcal{I}_\theta \big\} \tag{5.41}$$

Values of $\overline{Q}_\theta$ can be calculated by linear programms. To this end, choose any $x_k \in A_k$ for every $k \in \{1, \dots, r\}$. Then:
For any $f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$, consider

$$\big(f(x_1)\,,\;\dots\;,f(x_r)\big) \cdot q \;\;\rightarrow\;\; \max$$

where

$$\big(f_i(x_1)\,,\;\dots\;,f_i(x_r)\big) \cdot q \;\;\le\;\; \overline{P}'_\theta[f_i] + \varepsilon_i d_i \quad \forall\, i \in \mathcal{I}_\theta$$

and

$$q \in \mathbb{R}^r\,, \qquad q_k \ge 0 \quad \forall\, k \in \{1, \dots, r\}\,, \qquad q_1 + \,\dots\, + q_r \;=\; 1$$

The optimal value of this linear program is equal to $\overline{Q}_\theta[f]$.

<u>*STEP 3:*</u> Instead of the original imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ on the (infinite) sample space $(\mathcal{X}, \mathcal{A}')$, consider the imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ on the *finite* sample space $(\mathcal{X}, \mathcal{A})$ and solve the corresponding decision problem.

The following notation is used:

Notation 5.13 *Let $(\mathbb{D}, \mathcal{D})$ be a fixed decision space and let W be a fixed loss function. At first, $(\overline{P}'_\theta)_{\theta\in\Theta}$ is our imprecise model on the sample space $(\mathcal{X}, \mathcal{A}')$. The task is to find a randomized decision function on $(\mathcal{X}, \mathcal{A}')$ which is optimal*[11] *over all randomized decision functions on $(\mathcal{X}, \mathcal{A}')$. This decision problem is called* original decision problem.
Let the index set Θ, the decision space $(\mathbb{D}, \mathcal{D})$ and the loss function W remain unchanged. But, now, let the imprecise model be $(\overline{Q}_\theta)_{\theta\in\Theta}$ on the finite *sample space $(\mathcal{X}, \mathcal{A})$ where $(\overline{Q}_\theta)_{\theta\in\Theta}$ and $\mathcal{A}$ are constructed by the above discretization procedure. Then, the task is to find a randomized decision function on $(\mathcal{X}, \mathcal{A})$ which is optimal over all randomized decision functions on $(\mathcal{X}, \mathcal{A})$. This decision problem is called* (ε–)discretized decision problem.

[11] Here, the word "optimal" depends on the chosen optimization criterion such as Γ-minimaxity, E-admissibility, …

Theoretical properties of the discretization method

The theoretical properties of the discretization method presented above are summarized in the following theorem. It implies that the original decision problem may in fact be approximately solved by the discretized decision problem. For the optimization criterion Γ-minimaxity , this is explicated in Corollary 5.15 below.

Theorem 5.14 *Consider the setup of the present subsection. Then:*

a) For every randomized decision function τ on $(\mathcal{X},\mathcal{A})$, the risk function with respect to the discretized decision problem is approximately equal to the risk function with respect to the original decision problem – more precisely:

$$\begin{aligned} \sup_{Q_\theta \in \mathcal{N}_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau_x(dt)\, Q_\theta(dx) \;-\; \varepsilon(\sup W_\theta - \inf W_\theta) \;\leq& \\ \leq\;\; \sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau_x(dt)\, P'_\theta(dx) \;\;\leq& \\ \leq\;\; \sup_{Q_\theta \in \mathcal{N}_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau_x(dt)\, Q_\theta(dx)& \end{aligned}$$

for every $\theta \in \Theta$. Especially, the risk function with respect to the discretized decision problem is an upper bound for the risk function with respect to the original decision problem.

b) For every randomized decision function τ' on $(\mathcal{X},\mathcal{A}')$, there is a randomized decision function τ on $(\mathcal{X},\mathcal{A})$ such that the risk functions of τ and τ' satisfy

$$\begin{aligned} \sup_{Q_\theta \in \mathcal{N}_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau_x(dt)\, Q_\theta(dx) \;\;\leq& \\ \leq\;\; \sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\mathbb{D}} W_\theta(t)\, \tau'_x(dt)\, P'_\theta(dx) \;+\; \varepsilon(\sup W_\theta - \inf W_\theta)& \end{aligned}$$

for every $\theta \in \Theta$.

Proof: For every $\theta \in \Theta$, let $\overline{Q}'_\theta$ be the natural extension of $\overline{Q}_\theta$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. Accordingly, $\mathcal{N}'_\theta$ denotes the credal set of $\overline{Q}'_\theta$ on $(\mathcal{X},\mathcal{A}')$.

(a) Take any $\theta \in \Theta$ and recall the definitions in *STEP 1* of the proceeding of the discretization.

Then, for every $P'_\theta \in \mathcal{M}'_\theta$, the definition of s_i implies

$$\begin{aligned} P'_\theta[s_i] \;\leq\; P'_\theta\Big[f_i + \varepsilon_i d_i\Big] \;=\; P'_\theta[f_i] + \varepsilon_i d_i \;\leq& \\ \leq\; \overline{P}'_\theta[f_i] + \varepsilon_i d_i& \end{aligned}$$

for every $i \in \mathcal{I}_\theta$ and, therefore, the definition of $\overline{Q}'_\theta$ implies $P'_\theta \in \mathcal{N}'_\theta$. Hence,

$$\overline{P}'_\theta[f'] \;\leq\; \overline{Q}'_\theta[f'] \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \tag{5.42}$$

Next, consider the coherent upper prevision $\overline{Q}'_0$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ defined by

$$\overline{Q}'_0[f'] \;=\; \sup\left\{ \; Q'_0[f'] \;\middle|\; \begin{array}{c} Q'_0 \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')) \,, \\ Q'_0[f_i] \leq \overline{Q}'_\theta[f_i] \quad \forall\, i \in \mathcal{I}_\theta \end{array} \right\}$$

for every $f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. Together with (5.42), we have

$$\overline{P}'_\theta[f'] \;\le\; \overline{Q}'_\theta[f'] \;\le\; \overline{Q}'_0[f'] \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \tag{5.43}$$

and

$$\overline{Q}'_0[f_i] \;=\; \overline{Q}'_\theta[f_i] \qquad \forall\, i \in \mathcal{I}_\theta \tag{5.44}$$

Next, it follows from the definition of d_i that

$$\begin{aligned}\overline{P}'_\theta[f_i] \;&\le\; \overline{Q}'_0[f_i] \;\overset{(5.44)}{=}\; \overline{Q}'_\theta[f_i] \;\le\; \overline{Q}'_\theta[s_i] \;\le\; \overline{P}'_\theta[f_i] + \varepsilon_i d_i \;\le \\ &\le\; \overline{P}'_\theta[f_i] + \varepsilon_i\big(\overline{P}'_\theta[f_i] - \underline{P}'_\theta[f_i]\big) \qquad \forall\, i \in \mathcal{I}_\theta\end{aligned}$$

and from an application of Proposition 5.1 for $\overline{P}'_\theta$ and $\overline{Q}'_0$ that

$$\overline{P}'_\theta[f'] \;\le\; \overline{Q}'_0[f'] \;\le\; \overline{P}'_\theta[f'] + \varepsilon\big(\sup f' - \inf f'\big) \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

because $\overline{Q}'_0$ is the natural extension of a corresponding coherent upper prevision on $\mathcal{K}_\theta$ and the definitions ensure $\sum_{i\in\mathcal{I}_\theta} \varepsilon_i \le \varepsilon$. Hence, (5.43) implies

$$\overline{P}'_\theta[f'] \;\le\; \overline{Q}'_\theta[f'] \;\le\; \overline{P}'_\theta[f'] + \varepsilon\big(\sup f' - \inf f'\big) \tag{5.45}$$

for every $f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. Finally, part a) follows from (5.45).

(b) Let τ' be a randomized decision function on $(\mathcal{X},\mathcal{A}')$. Then, an application of Theorem 5.8 for the imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A})$ and its natural extension $(\overline{Q}'_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A}')$ implies the existence of a randomized decision function τ on $(\mathcal{X},\mathcal{A})$ such that

$$\sup_{Q_\theta\in\mathcal{N}_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tau_x(dt)\,Q_\theta(dx) \;\le\; \sup_{Q'_\theta\in\mathcal{N}'_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tau'_x(dt)\,Q'_\theta(dx)$$

for every $\theta \in \Theta$. Then, it follows that

$$\begin{aligned}&\sup_{Q_\theta\in\mathcal{N}_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tau_x(dt)\,Q_\theta(dx) \;\le \\ &\le\; \sup_{Q'_\theta\in\mathcal{N}'_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tau'_x(dt)\,Q'_\theta(dx) \;\le \\ &\overset{(5.45)}{\le}\; \sup_{P'_\theta\in\mathcal{M}'_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tau'_x(dt)\,P'_\theta(dx) \;+\; \varepsilon(\sup W_\theta - \inf W_\theta)\end{aligned}$$

for every $\theta \in \theta$.

□

The following corollary states that an approximately Γ - minimax randomized decision function in the original decision problem can be found by searching for a Γ - minimax randomized decision function in the discretized decision problem. Here, $\mathcal{T}_0(\mathcal{X},\mathbb{D})$ denotes the set of all randomized decision functions on $(\mathcal{X},\mathcal{A})$ and $\mathcal{T}'_0(\mathcal{X},\mathbb{D})$ denotes the set of all randomized decision functions on $(\mathcal{X},\mathcal{A}')$.

Corollary 5.15 *In the setup of the present subsection, let* $\overline{\Pi}$ *be a coherent upper prevision on* $\mathcal{L}_\infty(\Theta, 2^\Theta)$ *with credal set* $\mathcal{P}$ *. Let* $\tilde{\tau}$ *minimize the upper Bayes risk in the discretized decision problem, i.e.*

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta}, \tilde{\tau}, W\big) \;=\; \inf_{\tau\in\mathcal{T}_0(\mathcal{X},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta}, \tau, W\big)$$

Then, $\tilde{\tau}$ *is a randomized decision function on* $(\mathcal{X}, \mathcal{A}')$ *and approximately minimizes the upper Bayes risk in the original decision problem, i.e.*

$$R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tilde{\tau}, W\big) \;\le\; \inf_{\tau'\in\mathcal{T}'_0(\mathcal{X},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tau', W\big) \;+\; \varepsilon(\sup W - \inf W)$$

Proof: Take any $\hat{\tau}' \in \mathcal{T}'_0(\mathcal{X},\mathbb{D})$.

According to Theorem 5.14 b), there is some $\hat{\tau} \in \mathcal{T}_0(\mathcal{X},\mathbb{D})$ such that

$$\begin{aligned}\sup_{Q_\theta\in\mathcal{N}_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\hat{\tau}_x(dt)\,Q_\theta(dx) \;\;&\le \\ \le\;\; \sup_{P'_\theta\in\mathcal{M}'_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\hat{\tau}'_x(dt)\,P'_\theta(dx) \;&+\; \varepsilon(\sup W_\theta - \inf W_\theta)\end{aligned} \tag{5.46}$$

for every $\theta \in \Theta$. Hence, the definition of the upper Bayes risk (Section 3.2) implies

$$\begin{aligned}R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta}, \hat{\tau}, W\big) \;\;&\le \\ \le\;\; R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \hat{\tau}', W\big) \;&+\; \varepsilon(\sup W - \inf W)\end{aligned} \tag{5.47}$$

By use of Theorem 5.14 a) , it follows that

$$\sup_{P'_\theta\in\mathcal{M}'_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tilde{\tau}_x(dt)\,P'_\theta(dx) \;\;\le\;\; \sup_{Q_\theta\in\mathcal{N}_\theta} \int_{\mathcal{X}}\int_{\mathbb{D}} W_\theta(t)\,\tilde{\tau}_x(dt)\,Q_\theta(dx)$$

and, therefore,

$$R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tilde{\tau}, W\big) \;\le\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_{\theta\in\Theta}, \tilde{\tau}, W\big) \tag{5.48}$$

Assertions (5.47) and (5.48) and optimality of $\tilde{\tau}$ imply

$$\begin{aligned}R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \tilde{\tau}, W\big) \;\;&\le \\ \le\;\; R_{\overline{\Pi}}\big((\overline{P}'_\theta)_{\theta\in\Theta}, \hat{\tau}', W\big) \;&+\; \varepsilon(\sup W - \inf W)\end{aligned} \tag{5.49}$$

This proves Corollary 5.15 because (5.49) is true for every $\hat{\tau}' \in \mathcal{T}'_0(\mathcal{X},\mathbb{D})$. □

Though Troffaes (2008) is concerned with discretizing Θ, the setup of the present subsection is closely related to the setup in Troffaes (2008).

In the above described discretization method, the discrete sample space $(\mathcal{X},\mathcal{A})$ is generated by some simple functions s where every simple function s corresponds to some $f \in \mathcal{K}_\theta$ such that

$$\sup_{x\in\mathcal{X}} |f(x) - s(x)| \;=\; \max_{A\in\mathcal{A}} \sup_{x\in A} |f(x) - s(x)| \;\le\; \varepsilon(\sup f - \inf f)$$

This is denoted by

$$f \ \sim_\varepsilon \ s$$

in Troffaes (2008). Furthermore, $(\overline{Q}_\theta)_{\theta\in\Theta}$ is an imprecise model on $(\mathcal{X},\mathcal{A})$ where $(\mathcal{N}_\theta)_{\theta\in\Theta}$ is the corresponding family of credal sets on $(\mathcal{X},\mathcal{A})$. It can be read off from the construction of $(\overline{Q})_{\theta\in\Theta}$ and from the proof of Proposition 5.1 that

$$\inf_{P'_\theta\in\mathcal{M}'_\theta} \|Q_\theta - P'_\theta\| \ \le \ 2\varepsilon \qquad \forall\, Q_\theta \in \mathcal{N}_\theta$$

and

$$\inf_{Q_\theta\in\mathcal{N}_\theta} \|P'_\theta - Q_\theta\| \ \le \ 2\varepsilon \qquad \forall\, P'_\theta \in \mathcal{M}'_\theta$$

for every $\theta \in \Theta$. [12] [13] This is denoted by

$$\mathcal{M}'_\theta \ \sim_{2\varepsilon} \ \mathcal{N}_\theta$$

in Troffaes (2008). Furthermore, adopting the terminology from Troffaes (2008), Corollary 5.15 may be reformulated in the following way:

> *Every randomized decision function on $(\mathcal{X},\mathcal{A})$ which is optimal in the discretized decision problem is ε-optimal in the original decision problem.*

Accordingly, Corollary 5.15 corresponds to (Troffaes, 2008, Theorem 6). However, note that Corollary 5.15 is concerned with discretizing the sample space $(\mathcal{X},\mathcal{A}')$ whereas (Troffaes, 2008, Theorem 6) is concerned with discretizing Θ .

Applicability of the discretization method

The above presented discretization method can be applied step by step. Especially, every value which has to be calculated can in principle be calculated by linear programming. However, rigid applications may in general be handicapped – or even made impossible – because of exceedingly high computational costs. This is again similar to the results in Troffaes (2008) and we may derive upper bounds for the size of the discretized sample

[12] $\|\cdot\|$ denotes the operator norm in $\mathrm{ba}_1^+(\mathcal{X},\mathcal{A})$, i.e.

$$\|Q_\theta - P'_\theta\| \ = \ \sup_{f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})} \frac{|Q_\theta[f] - P'_\theta[f]|}{\|f\|}$$

[13] In addition, this is a direct consequence of the following fact:

> Let $\overline{P}_1$ and $\overline{P}_2$ be coherent upper previsions on $(\mathcal{X},\mathcal{A})$ with credals sets $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Then,
>
> $$\sup_{f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})} \frac{|\overline{P}_1[f] - \overline{P}_2[f]|}{\|f\|} \ = \ \max\{\delta_1\,,\ \delta_2\}$$
>
> where
>
> $$\delta_1 \ = \ \sup_{P_1\in\mathcal{M}_1}\ \inf_{P_2\in\mathcal{M}_2} \|P_1 - P_2\| \qquad \text{and} \qquad \delta_2 \ = \ \sup_{P_2\in\mathcal{M}_2}\ \inf_{P_1\in\mathcal{M}_1} \|P_2 - P_1\|$$

The proof of this fact arose from a discussion of Damjan Skulj and the author at the *Workshop on Principles and Methods of Statistical Inference with Interval Probability*, Durham, 12-16 May 2008. A publication containing this proof will follow.

space which generally holds but which are, in general, much too large in order to be of any practical value:

As stated before, there is a finite partition $\{A_1, \ldots, A_r\}$ of $\mathcal{A}$ such that every element of $\mathcal{A}$ is the union of some elements of the partition $\{A_1, \ldots, A_r\}$. The size of this partition – i.e. the number $r \in \mathbb{N}$ – precisely corresponds to the size of the discretized sample space: r is the number of possible (discrete) observations after discretizing.
According to the definition of $\mathcal{A}$, the partition $\{A_1, \ldots, A_r\}$ is the coarsest partition which is finer than every partition

$$\{A_i^{(1)}, \ldots, A_i^{(M)}\}, \qquad i \in \{1, \ldots, n\}$$

where $A_i^{(j)}$ is defined in (5.38) for every $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, M\}$.
Therefore, an upper bound on r is given by

$$r \;\leq\; M^n \;\leq\; \left(1 + \frac{1}{\varepsilon} \cdot \sup_{\theta \in \Theta} \sum_{j \in \mathcal{I}_\theta} \frac{\sup f_j - \inf f_j}{d_j}\right)^n \tag{5.50}$$

where the last inequality follows from (5.36) and (5.37).
This number is extremely large – even if Θ is a small set and, for every $\theta \in \Theta$, $\mathcal{K}_\theta$ only contains a few elements. For example, let Θ contain 10 elements, and, for every $\theta \in \Theta$, let each $\mathcal{K}_\theta$ also contain 10 elements such that $\mathcal{K}_{\theta_1}, \ldots, \mathcal{K}_{\theta_{10}}$ are pairwise disjoint. Therefore, we have $n = 100$. Furthermore, assume for simplicity that

$$\overline{P}'_\theta[f] - \underline{P}'_\theta[f] \;=\; 0.1 \cdot (\sup f - \inf f) \qquad \forall\, f \in \mathcal{K}_\theta \qquad \forall\, \theta \in \Theta$$

Then, for $\varepsilon = 0.1$, the number in (5.50) is

$$\left(1 + \tfrac{1}{0.1} \cdot 10 \cdot \tfrac{1}{0.1}\right)^{100} \;>\; 10^{300}$$

However, this number usually decreases immensely: It is unrealistic to assume that $\mathcal{K}_{\theta_1}, \ldots, \mathcal{K}_{\theta_{10}}$ are pairwise disjoint in applications. In most applications, $\mathcal{K}_\theta$ will not depend on θ so that we have

$$\mathcal{K} \;=\; \mathcal{K}_\theta \qquad \forall\, \theta \in \Theta$$

In this case, n does not increase with the number of elements of Θ and we would get $n = 10$ in the above example. This leads to the number

$$\left(1 + \tfrac{1}{0.1} \cdot 10 \cdot \tfrac{1}{0.1}\right)^{10} \;\approx\; 10^{30}$$

which still is a great deal too large. However, (5.50) only is a very crude upper bound which does not assume any additional properties of the functions $f \in \mathcal{K}$. Such assumptions may drastically decrease the bound as can be seen by Proposition 5.16.

Proposition 5.16 *Let $\mathcal{X}$ be an interval in $\mathbb{R}$ and assume that every $f \in \mathcal{K}$ fulfills one of the following properties:*

(a) f is the indicator function of a set $A' \in \mathcal{A}'$ which is the union of no more than

$$1 + \frac{1}{\varepsilon} \cdot \sup_{\theta \in \Theta} \sum_{j \in \mathcal{I}_\theta} \frac{\sup f_j - \inf f_j}{d_j}$$

intervals

(b) f is convex.

(c) f is concave.

Let r be the number of elements of the partition $\{A_1,\dots,A_r\}$. Then,

$$r \;\leq\; 4n\cdot\left(1+\frac{1}{\varepsilon}\cdot\sup_{\theta\in\Theta}\sum_{j\in\mathcal{I}_\theta}\frac{\sup f_j-\inf f_j}{d_j}\right) \tag{5.51}$$

Proof: Recall that $\mathcal{K}=\{f_1,\dots,f_n\}$ and recall that $\{A_i^{(1)},\dots,A_i^{(M)}\}$ is the partition defined by (5.38) for every $i\in\{1,\dots,n\}$.

Firstly, it is shown for every $i\in\{1,\dots,n\}$ that there is a partition $\{C_i^{(1)},\dots,C_i^{(2M)}\}$ of $\mathcal{X}$ such that

- $C_i^{(j)}$ is an intervall in $\mathbb{R}$ for every $j\in\{1,\dots,2M\}$ and
- the smallest σ-algebra generated by $\{C_i^{(1)},\dots,C_i^{(2M)}\}$ contains $\{A_i^{(1)},\dots,A_i^{(M)}\}$.

For any $i\in\{1,\dots,n\}$ such that f_i fulfills (a), this follows immediately from (5.37). Now, take any $i\in\{1,\dots,n\}$ such that f_i fulfills (b). Then, the definition of $A_i^{(j)}$ and convexity of f_i implies that $A_i^{(j)}$ is the union of two intervals $C_i^{(j)}$ and $C_i^{(2j)}$ for every $j\in\{1,\dots,M\}$. The same is true in case of (c).

Next, note that the number of elements of each partition $\{C_i^{(1)},\dots,C_i^{(2M)}\}$ is bounded by

$$2M \;\leq\; 2\cdot\left(1+\frac{1}{\varepsilon}\cdot\sup_{\theta\in\Theta}\sum_{j\in\mathcal{I}_\theta}\frac{\sup f_j-\inf f_j}{d_j}\right)$$

Finally, Proposition 5.16 follows from the following simple fact:
Let $D_1^{(1)},\dots,D_1^{(m_1)}$ and $D_2^{(1)},\dots,D_2^{(m_2)}$ be two partitions of an interval in $\mathbb{R}$ such that every $D_i^{(j)}$ is an intervall. Then, there is a partition $D_1,\dots,D_{r'}$ such that every $D_i^{(j)}$ is the union of some elements of $\{D_1,\dots,D_{r'}\}$ and the size r' of this partition is not larger than $m_1+2\cdot m_2$. □

In the situation of the above example with $n=10$, this leads to the upper bound

$$4\cdot 10\cdot\left(1+\tfrac{1}{0.1}\cdot 10\cdot\tfrac{1}{0.1}\right)\;\approx\;4\cdot 10^4$$

which is a more reasonable size than the above ones. In particular, bound (5.51) has the remarkable property that it increases only linearly(!) in n, the number of functions. On the one hand, Proposition 5.16 itself covers many situations in real applications. On the other hand, it demonstrates, that applying the presented discretization procedure will often lead to a reasonable size r of the discretized sample space.

Furthermore, there is another way to reduce the size of the discretized sample space $(\mathcal{X},\mathcal{A})$, which is different from the others and relates to the results of Section 5.2. In the presented discretization method, an imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ is constructed such that

$$\overline{P}'_\theta[f_i]\;\leq\;\overline{Q}_\theta[f_i]\;\leq\;\overline{P}'_\theta[f_i]+\frac{\varepsilon}{c}(\sup f_i-\inf f_i)\qquad\forall\, i\in\{1,\dots,n\} \tag{5.52}$$

However, c can be large and then, it will not be possible in most applications to specify the "correct" coherent upper prevision in such a great precision that the value

$$\varepsilon_i d_i \;=\; \frac{\varepsilon}{c}(\sup f_i - \inf f_i)$$

becomes meaningful in (5.52).
Therefore, it may be justified to relax (5.52) to

$$\overline{P}'_\theta[f_i] \;\leq\; \overline{Q}_\theta[f_i] \;\leq\; \overline{P}'_\theta[f_i] + \varepsilon(\sup f_i - \inf f_i) \qquad \forall\, i \in \{1,\ldots,n\} \tag{5.53}$$

This means, that M is not chosen in order to fulfill (5.37) in the discretization method. Instead, M has to be chosen so that

$$M - 1 \;<\; \frac{1}{\varepsilon} \;\leq\; M$$

Then, analog to (5.51), an upper bound on the size r would be

$$4 \cdot n \cdot \left(1 + \frac{1}{\varepsilon}\right) \tag{5.54}$$

and the above example would lead to

$$4 \cdot 10 \cdot \left(1 + \tfrac{1}{0.1}\right) \;=\; 440$$

This is a reasonable size with which computations should be tractable. Note that bound (5.54) does not depend on the size of Θ and only depends linearly on the number of elements in $\mathcal{K}$. Therefore, also larger problem than the above example should be tractable. Relaxing (5.52) to (5.53) can often be justified and leads to more conservative results. However, note that, by doing this, ε-optimality is not guaranteed anymore according to the results in Section 5.2.

Chapter 6

Application: Minimum distance estimation

6.1 Introduction

The present chapter is concerned with an application of statistical decision theory – namely estimating, which is, in addition to hypothesis testing, one of the most important issues in statistics. While hypothesis testing under imprecise probabilities has been extensively studied – especially by T. Augustin in (Augustin, 1998) and (Augustin, 2002) on base of the Huber-Strassen theory [1], estimating a parameter has hardly been considered explicitly within the theory of coherent upper previsions so far. There are a few articles which are concerned with it in Bayesian models (primarily associated with Walley's Imprecise Dirichlet Model), e.g. Walley (1996), Quaeghebeur and de Cooman (2005), Hutter (2008) and Walter and Augustin (2008). In addition, there are a few articles which address very special applications, e.g. Kriegler and Held (2003) (climate projections) and Bickis and Bickis (2007) (prediction of the next influenza pandemic). However, there does not seem to be any publication which is concerned with general frequentist estimation of a parameter using coherent upper/lower previsions. [2] Therefore, the present chapter cannot be restricted to the sole investigation of the proposed minimum distance estimator but also has to develop some fundamentals of (frequentist) estimating under coherent upper previsions at first. This is necessary the more so as the minimum distance estimator is associated with the empirical process (which needs a somewhat more elaborated setting) and is justified by asymptotic arguments (but an elaborated asymptotic theory of imprecise probabilities is still missing).

In the spirit of Wald (1950), an estimation problem may be defined in the following way: Take the decision space to be equal to

$$(\mathbb{D}, \mathcal{D}) \;=\; (\Theta, 2^{\Theta})$$

and let

$$W \;:\; \Theta \times \Theta \;\to\; \mathbb{R}\,, \qquad (\theta, \hat{\theta}) \;\mapsto\; W_\theta(\hat{\theta})$$

[1] see also Augustin (2002) for a review of the work following Huber and Strassen (1973)

[2] In a somewhat different setting, Cozman and Chrisman (1997), Fierens and Fine (2003) and Rêgo and Fine (2005) also consider estimation problems where coherent upper previsions are interpreted in an objective, frequentist way. However, they do not consider the estimation of a parameter in an imprecise model $(\overline{P}_\theta)_{\theta\in\Theta}$ but the estimation of a totally unknown credal set $\mathcal{M}$.

be a loss function such that $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\Theta, 2^\Theta)$. Then, a decision $t = \hat{\theta} \in \Theta = \mathbb{D}$ is called *estimation of the true parameter* θ ; (randomized) decision functions on $(\mathcal{X}, \mathcal{A}')$ are called *estimators* and the performance of estimators is evaluated by their risk functions.

Though statistical evaluations are usually not based on one single observation x but on several or even many observations $y_1, \ldots, y_n$, this is also covered by the above formalization. In mathematical statistics, all observations are commonly treated as one observation by putting

$$x \;=\; (y_1, \ldots, y_n)$$

Accordingly, it is assumed that $\mathcal{X}$ is a suitable product space

$$\mathcal{X} \;=\; \mathcal{Y}^n\,, \qquad y_i \in \mathcal{Y} \qquad \forall\, i \in \{1, \ldots, n\}$$

and that the (precise) distribution of the data is a product probability measure

$$P'_\theta \;=\; Q_\theta^{\otimes n}$$

In this way, it is often possible to treat several observations as one single observation. Therefore, it is enough without loss of generality to take only one single observation into account. An analog proceeding is also possible in case of imprecise probabilities:
To this end, let the coherent upper prevision $\overline{Q}_\theta$ be the imprecise distribution of the observation $y_i \in \mathcal{Y}$ and let $\mathcal{N}_\theta$ denote its credal set. Then, there are several different possibilities of defining independent products of coherent upper previsions $\overline{Q}_\theta$. One of the most common definition of such a product is the so-called *type-1 product.* According to this definition given by (Walley, 1991, § 9.3.5), the product is the coherent upper prevision $\overline{P}'_\theta$ which corresponds to the credal set

$$\mathcal{M}'_\theta \;=\; \mathit{clco}\big\{\, Q_1 \otimes \cdots \otimes Q_n \;\big|\; Q_i \in \mathcal{N}_\theta \quad \forall\, i \in \{1, \ldots, n\}\big\} \tag{6.1}$$

By doing this, estimation problems under imprecise probabilities may simply be treated as special cases of data-based decision theory under imprecise probabilities. This is true from a theoretical point of view but severe problems will arise in applications because the complexity of the above credal set $\mathcal{M}'_\theta$ drastically increases with the number of observations. Therefore, $\mathcal{M}'_\theta$ may be computational tractable for $n = 10$ but it will rarely be tractable for $n = 50$ or larger numbers of observations. In case of hypothesis testing, this has already been discussed in (Augustin, 1998, § 4.1.4 and § 6.1.2). At least in case of hypothesis testing, this problem can sometimes be avoided: In the presence of *globally* least favorable pairs, the complexity of the testing problem does not increase with the number of observations; cf. (Rieder, 1974, Satz II.B.4), (Witting, 1985, Satz 2.57) and (Augustin, 1998, § 6.1.2). Least favorable pairs have attracted much attention after the publication of Huber and Strassen (1973). However, it has also been shown in Huber and Strassen (1973) that globally least favorable pairs only exist in case of two-alternating capacities which are very special cases of imprecise probabilities. Confer Augustin (1998) and Augustin (2002) for a detailed review (including many references) of this so-called "Huber-Strassen theory". In case of more general imprecise probabilities, Augustin (1998) considers *local* least favorable pairs.[3] Though this local concept is similar to the global

[3] The least favorable models considered in Chapter 4 are a generalization of these *local* least favorable pairs.

one, it is not known if and in how far it can be used in order to reduce the increase of the computational effort relating to the type-1 product

$$\mathcal{M}'_\theta \;=\; \mathit{clco}\big\{\, Q_1 \otimes \cdots \otimes Q_n \;\big|\; Q_i \in \mathcal{N}_\theta \quad \forall\, i \in \{1,\ldots,n\}\big\}$$

if the sample size n increases. [4] [5]

Instead of the type-1 product, the type-2 product is used in the following. Here, $\mathcal{X}$ is not assumed to be a product space $\mathcal{X} = \mathcal{Y}^n$. Accordingly, x does not represent all observations but x is really only one single observation which is distributed according to some coherent upper prevision $\overline{P}'_\theta$. If we have a number of such observations

$$x_1 \in \mathcal{X}\,,\;\ldots\,,\; x_n \in \mathcal{X}$$

we have to consider the product space $\mathcal{X}^n$. Using the type-2 product corresponds to the assumption that the vector $(x_1,\ldots,x_n)$ containing all observations is distributed according to some

$$P'_\theta \otimes \cdots \otimes P'_\theta\,, \qquad P'_\theta \in \mathcal{M}'_\theta$$

where $\mathcal{M}'_\theta$ denotes the credal set of $\overline{P}'_\theta$.[6] Imprecise probabilities may be interpreted in different ways but note that the type-2 product is only compatible with the interpretation of sensitivity analysts. Accordingly, the type-2 product is commonly used in robust statistics.

The definition of the type-2 product is recalled and reformulated in terms of random variables in Subsection 6.2.1. Such a reformulation is necessary because the proposed minimum distance estimator is associated with the empirical process. Therefore, the investigations need a rather elaborated setting which is based on random variables and image measures. These are fundamental concepts in classical statistics which, initially,

[4] See (Augustin, 1998, p. 238ff).

[5] In addition to this problem, the use of the type-1 product in estimation problems is also handicapped by some more issues:

A lot of current research in statistics is based on simulations. Though simulations are completely useless in order to assure good properties of statistical methods, they can often provide valuable insight and intuition into the behavior of the statistical method in real applications. However, it is not clear what simulating means within the theory of imprecise probabilities. Especially, this seems to apply to the above product model. Theoretical investigations and a practicable guideline for "good" simulations are still missing at the moment. Such simulations should be able to broadly cover the manifold aspects of the above product model – otherwise they will commonly fail at providing valuable insight. The meaning of simulating within the theory of imprecise probabilities has also been discussed in a session at the *Workshop on Principles and Methods of Statistical Inference with Interval Probability*, Durham, 12-16 May 2008.

Next, asymptotic properties are also important criteria for the evaluation of statistical methods but an asymptotic theory of imprecise probabilities which applies to the above product model has not been developed yet.

Finally, at least for sensitivity analysts, the so-called type-2 product seems to be a better approximation of the real world in most applications than the type-1 product.

Of course, the latter point depends on the situation and touches philosophic discussions. Note that the other points do not argue against type-1 products in general! Instead, they are intended to indicate that a lot of research has to be done before statistical methods based on type-1 products can be developed, proven to have good theoretical properties and applied in real applications with reasonable numbers of observations. However, these things seem to be already possible for type-2 products – which is the purpose of the following Sections.

[6] Cf. e.g. (Walley, 1991, § 9.3.5) and (Couso et al., 1999, § 3.6).

should be carried over to statistics with imprecise probabilities. After this, Subsection 6.2.2 is concerned with an ad hoc method of discretizing the parameter space Θ which may serve as a "less-than-ideal solution" if modeling yields an inifinte parameter space Θ – confer Subsection 5.4.1.

The minimum distance estimator is defined in Section 6.3. In short, the estimator is based on the following simple idea: The data $x_1, \ldots, x_n$ – which are independent identically distributed according to a coherent upper prevision $\overline{P}'_\theta$ – are used to build the empirical measure

$$\mathbb{P}^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

Then, the minimum distance estimator is that $\hat{\theta} \in \Theta$ such that $\mathbb{P}^{(n)}$ lies next to $\mathcal{M}'_{\hat{\theta}}$ where $\mathcal{M}'_\theta$ denotes the credal set of $\overline{P}'_\theta$ for every $\theta \in \Theta$.

Though such minimum distance estimators can be defined for any coherent upper previsions, the investigations in the present chapter are based on the following crucial assumption on the credal sets: It is assumed that every credal set is given by

$$\mathcal{M}'_\theta \;=\; \left\{P'_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \;\middle|\; P'_\theta[f] \le \overline{P}'_\theta[f] \quad \forall\, f \in \mathcal{K}\right\}, \qquad \theta \in \Theta$$

where $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ is a finite subset of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ which does not depend on θ.
On the one hand, such credal sets are important in practical applications (cf. p. 161). On the other hand, this assumption guarantees good asymptotic properties of the estimator (Section 6.4) and enables an efficient implementation (Section 6.5).
In fact, the asymptotic properties turn out to be even better than in case of precise probabilities: Though the total variation norm is used, it is shown in Section 6.4 that – based on this norm – the distance between the empirical measure and the correct imprecise probability converges to zero. This is not necessarily true for precise probabilities. Furthermore, it is shown that the rate of convergence is at least of order

$$O\!\left(\frac{\ln\ln n}{\sqrt{n}}\right)$$

– a rate which is known in connection with the strong law of large numbers for precise probabilities. It follows from these results that the proposed minimum distance estimator is consistent.
Next, Section 6.5 is concerned with the implementation of the estimator. In Subsections 6.5.1 and 6.5.2, an algorithm for calculating distances between the empirical measure and coherent upper previsions is developed and its correctness is proven. These evaluations rely on the results obtained in Chapter 5. It is shown, that – after a suitable discretization – the distance can be calculated by a linear program. Fortunately, this linear program only modestly increases with the number of observations and, as a consequence, the minimum distance estimator can also be calculated for many observations. E.g. in Section 6.6, the estimator is applied in a simulation study with 500 runs and 10000 observations in every run. The estimator has been programmed in R and is already publicly available as R package "imprProbEst"; cf. Hable (2008a). Section 6.6 presents applications of the estimator in three different models and the numerical results demonstrate that the minimum distance estimator is practicable (due to often exceedingly high computational costs, this is not self-evident within imprecise probabilities) and yields good results in comparison with estimators developed for precise probabilities.

The present chapter shows that it is possible to successfully work with the imprecise models (6.1) even though these models are quite extensive. In particular, these models are much more extensive than parametrically generated imprecise models. It is not investigated here in how far the complexity of models (6.1) can be reduced by applications of the concept of "sufficiency" defined in Subsection 3.3.2. Though it has been shown in Subsection 3.3.2.2 how sufficiency can be used in order to deal with parametrically generated imprecise models, more advanced applications would have been out of the scope of the present book as the definition of sufficiency only resulted as a welcomed byproduct from the investigations in Chapter 3.

Minimum distance estimators have already attracted attention in robust statistics [7] but our setup considerably differs from those ones usually used in robust frequentist estimation. In particular, the credal sets in (6.1) are not in accordance with common neighborhood systems. There, it is assumed that the observations are approximately distributed according to a *known* ideal precise model. As a consequence, concepts which are common in robust statistics (e.g. influence functions) cannot be applied offhand in the theory of imprecise probabilities.

6.2 Estimation in an imprecise probability model

6.2.1 Independent observations and random variables

In a classical estimation problem, we have a parametric family $(P'_\theta)_{\theta\in\Theta}$ of precise probability distributions on a sample space $(\mathcal{X},\mathcal{A}')$. The task is to estimate the true parameter $\theta_0\in\Theta$. Most often, it is assumed that the estimation can be based on a whole set of data

$$x_1\,,\;\ldots\;,\;x_n\;\;\in\;\;\mathcal{X}$$

which are independent identically distributed according to the true distribution P'_{θ_0} . That is, the vector $x=(x_1,\ldots,x_n)$ consisting of all observations is distributed according to the product measure ${P'_{\theta_0}}^{\otimes n}$.

In a (more realistic) imprecise probability setup, it is natural to replace the precise model $(P'_\theta)_{\theta\in\Theta}$ by an imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ which consists of coherent upper previsions $\overline{P}'_\theta$. Hence, it is assumed that the data

$$x_1\,,\;\ldots\;,\;x_n\;\;\in\;\;\mathcal{X}$$

are independent identically distributed according to the true $\overline{P}'_{\theta_0}$ or – in other words – the vector $x=(x_1,\ldots,x_n)$ consisting of all observations is distributed according to a coherent upper product prevision ${\overline{P}'_{\theta_0}}^{\otimes n}$.

As stated in the introductory Section 6.1, there are several different ways to define such products of coherent upper previsions. In the following, the *type-2 product*[8] is used which corresponds to a strict sensitivity analyst's point of view. This product prevision is defined to be that coherent upper prevision

$${\overline{P}'_\theta}^{\otimes n}\;:\;\mathcal{L}_\infty\big(\mathcal{X}^n,\mathcal{A}'^{\otimes n}\big)\;\;\to\;\;\mathbb{R}$$

[7] See e.g. Parr and Schucany (1980), Millar (1981), Donoho and Liu (1988), (Rieder, 1994, § 6) and Öztürk and Hettmansperger (1998)

[8] cf. (Walley, 1991, § 9.3.5)

which has credal set

$$c\ell\,\mathrm{co}\left\{\,P_\theta'^{\otimes n}\;\Big|\;P_\theta' \in \mathcal{M}_\theta'\,\right\}$$

where $\mathcal{M}_\theta'$ denotes the credal set of $\overline{P}_\theta'$.

Though this definition of the type-2 product is commonly used, it is not enough elaborated for the following investigations. This is because the minimum distance estimator is based on the empirical measure and, therefore, we have to deal with stochastic processes. In this context, a detailed mathematical formulation of the setup is necessary. This is done by use of random variables and image measures in classical probability theory and mathematical statistics. In the following, it is shown how this formalization can be adopted for imprecise probabilities.

Firstly, let us recall the classical setup: There, a random observation or data point x_0 in a set $\mathcal{X}$ is mathematically formalized by a map

$$X_0 \;:\; \Omega \;\to\; \mathcal{X}\,, \qquad \omega \;\to\; X_0(\omega)$$

where Ω is a fixed set which is rarely specified more closely. There are a fixed σ-algebra $\mathcal{F}$ on Ω and a fixed σ-algebra $\mathcal{A}'$ on $\mathcal{X}$ and it is assumed that X_0 is measurable with respect to these σ-algebras. X_0 is called *random variable.*

Next, it is assumed that an unspecified event ω has randomly happened which, by (deterministic) physical principles, has led to the observation

$$x_0 \;=\; X_0(\omega)$$

The events $\omega \in \Omega$ are distributed according to a (precise) distribution U or a (precise) distribution U_θ on $(\Omega, \mathcal{F})$ where θ is an unknown parameter.

Let $A' \in \mathcal{A}'$ be a measurable subset of $\mathcal{X}$. Then, the probability that the observation x_0 lies in A' is equal to

$$U_\theta\Big(\{\omega \in \Omega \,|\, X_0(\omega) \in A'\}\Big)$$

That is, x_0 is distributed according to the precise probability measure

$$P_\theta' \;:\; \mathcal{A}' \;\to\; [0,1]\,, \qquad A' \;\mapsto\; U_\theta\Big(\{\omega \in \Omega \,|\, X_0(\omega) \in A'\}\Big) \tag{6.2}$$

This defines a (precise) statistical model $(P_\theta')_{\theta\in\Theta}$ for the observation x_0 . P_θ' defined by (6.2) is called *image measure* of U_θ under X_0 and is denoted by $P_\theta' = X_0(U_\theta)$.

A whole set of observations/data $x_1, \dots, x_n$, is modeled via several random variables

$$X_i \;:\; \Omega \;\to\; \mathcal{X}\,, \qquad i \in \{1, \dots, n\}$$

Accordingly, it is assumed that the (unspecified) event $\omega \in \Omega$ has led to the observations/data

$$x_1 \;=\; X_1(\omega)\,,\; \dots\,,\; x_n \;=\; X_n(\omega)$$

The random variables

$$X_i \;:\; \Omega \;\to\; \mathcal{X}$$

are called *independent identically distributed* with respect to U_θ if their joint image measure is equal to the product of the single image measures and these image measures coincide:

$$\begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_n \end{pmatrix} \big(U_\theta\big) \;=\; X_1(U_\theta) \otimes \cdots \otimes X_n(U_\theta) \;=\; P_\theta'^{\otimes n}$$

Now, let us turn over to imprecise probabilities again: Due to our sensitivity analyst's point of view, it is assumed in the imprecise probability setup that there is a coherent upper prevision $\overline{U}_\theta$ and the distribution U_θ of the events $\omega \in \Omega$ is unknown and can be any element of the credal set $\mathcal{U}_\theta$ of $\overline{U}_\theta$.
Analogously to the ordinary image measure, we can define the image of a coherent upper prevision:

Definition 6.1 *The upper coherent prevision $\overline{P}_\theta'$ on $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ which corresponds to the credal set*

$$\mathcal{M}_\theta' \;=\; \big\{ X_0(U_\theta) \;\big|\; U_\theta \in \mathcal{U}_\theta \big\} \tag{6.3}$$

is called image of $\overline{U}_\theta$ under X *and is denoted by*

$$\overline{P}_\theta' \;=\; X(\overline{U}_\theta)$$

Lemma 6.2 below shows that this is defined well. That is, the image of a coherent upper prevision is again a coherent upper prevision. This provides a nice generalization of classical probability theory which is based on the fact that the image of a probability measure is again a probability measure.
In this way, we get an imprecise model $(\overline{P}_\theta')_{\theta\in\Theta}$. Since U_θ is any element of the credal set $\mathcal{U}_\theta$, the distribution of the observation x_0 modeled by the random variable X_0 is any element of the credal set $\mathcal{M}_\theta'$. The essential difference to the precise setting is the that, given θ, the true $U_\theta \in \mathcal{U}_\theta$ and, accordingly, the true $P_\theta' \in \mathcal{M}_\theta'$ are totally unknown.

Lemma 6.2 $\mathcal{M}_\theta$ *defined by (6.3) is a credal set on $(\mathcal{X}, \mathcal{A}')$.*

Proof: The map

$$\xi \;:\; \mathrm{ba}(\Omega, \mathcal{F}) \;\mapsto\; \mathrm{ba}(\mathcal{X}, \mathcal{A}')\,, \qquad \nu \;\mapsto\; \xi(\nu)$$

defined by

$$\xi(\nu)(A') \;=\; \nu\big(X_0^{-1}(A')\big)$$

is linear and continuous with respect to the $\mathcal{L}_\infty(\Omega, \mathcal{F})$ - topology on $\mathrm{ba}(\Omega, \mathcal{F})$ and the $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ - topology on $\mathrm{ba}(\mathcal{X}, \mathcal{A}')$. Together with

$$\xi\Big(\mathrm{ba}_1^+(\Omega, \mathcal{F})\Big) \;\subset\; \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}')$$

this implies that $\xi(\mathcal{U}_\theta)$ is a convex and $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ - compact subset of $\mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}')$. According to Corollary 2.16, $\xi(\mathcal{U}_\theta)$ is a credal set and the definitions imply

$$\xi(\mathcal{U}_\theta) \;=\; \mathcal{M}_\theta$$

□

Just as in the precise case, it is assumed that the random variables

$$X_i \;:\; \Omega \;\to\; \mathcal{X}_i\,, \qquad\quad i \in \{1,\dots,n\}$$

are independent identically distributed. That is, the joint distribution of observations is equal to

$$\begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_n \end{pmatrix}(U_\theta) \;\;=\;\; X_1(U_\theta)\otimes\cdots\otimes X_n(U_\theta) \;\;=\;\; P_\theta'^{\otimes n}$$

Since U_θ may be any element of $\mathcal{U}_\theta$, the distribution of the vector $x=(x_1,\dots,x_n)$ containing all observations may be any element of

$$\mathcal{N}_\theta' \;:=\; \left\{\, P_\theta'^{\otimes n} \;\middle|\; P_\theta' \in \mathcal{M}_\theta' \,\right\}$$

This set of product probabilities defines a coherent upper prevision

$$\overline{P}_\theta'^{\otimes n} \;:\; \mathcal{L}_\infty\big(\mathcal{X}^n, \mathcal{A}'^{\otimes n}\big) \;\;\to\;\; \mathbb{R}\,, \quad g' \;\;\mapsto\;\; \sup_{P_\theta'^{\otimes n}\in\mathcal{N}_\theta'} P_\theta'^{\otimes n}[g]$$

According to Proposition 2.15, the credal set of this coherent upper prevision is equal to

$$c\ell\,\mathrm{co}\left\{\, P_\theta'^{\otimes n} \;\middle|\; P_\theta' \in \mathcal{M}_\theta' \,\right\}$$

so that, in fact, we end up with the usual type-2 product of coherent upper previsions again.

Note that the credal sets $\mathcal{M}_\theta'$ may also contain probability charges which are not σ-additive. Products of probability charges such as $P_\theta'^{\otimes n}$ are defined according to (König, 1997, Proposition 20.4). However, these products are not defined on the product σ-algebra $\mathcal{A}'^{\otimes n}$ but on the (usually) smaller product algebra denoted by $\mathcal{A}'^{\hat{\otimes} n}$. This is the smallest algebra on $\mathcal{X}^n$ which contains all rectangles

$$A_1' \times \ldots \times A_n' \;\;\subset\;\; \mathcal{X}^n \qquad\text{where}\qquad A_1'\,,\;\ldots\,,\; A_n \;\;\in\;\; \mathcal{A}'$$

That is, $\overline{P}_\theta'^{\otimes n}$ is defined on the product algebra $\mathcal{A}'^{\hat{\otimes} n}$ at first. Next, $\overline{P}_\theta'^{\otimes n}$ can be extended to a coherent upper prevision on the usual product σ-algebra $\mathcal{A}'^{\otimes n}$ by natural extension.

6.2.2 Discretizations in estimation problems

As argued in Subsection 5.4.1, discretizing the parameter space Θ may be considered as part of modeling in estimation problems because coarsening Θ also means to change the purpose of the estimation problem and this change of the purpose is desirable from the point of view of the theory of imprecise probabilities; confer Subsection 5.4.1.

Modelers will nevertheless often produce an infinite parameter space Θ. Therefore, an ad hoc method for discretizing Θ is developed in the following:

Let Θ be any index set and $(\overline{P}_\theta')_{\theta\in\Theta}$ be an imprecise model on a sample space $(\mathcal{X},\mathcal{A}')$.
For every $\theta\in\Theta$, let $\mathcal{M}_\theta'$ be the credal set of $\overline{P}_\theta'$ on $(\mathcal{X},\mathcal{A}')$.
In order to discretize Θ, let

$$\mathcal{H} \;=\; \big\{H_1\,,\;\ldots\,,\; H_m\big\}$$

be a finite partition of Θ. Now, the parameter set in our estimation problem is $\mathcal{H}$ and we want to estimate the true $H \in \mathcal{H}$. This is the set $H \in \mathcal{H}$ in which the true parameter θ lies. That is, we do not want to discriminate between different elements θ_1 and θ_2 of one H any more. In this sense, the estimation problem gets coarser. The (upper) risk function depending on $H \in \mathcal{H}$ is canonically defined by

$$\mathcal{H} \;\to\; \mathbb{R}\,, \qquad H \;\mapsto\; \sup_{\theta \in H} \sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\Theta} W_\theta(\hat{\theta})\, \tau_x(d\hat{\theta})\, P'_\theta(dx) \tag{6.4}$$

where $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\Theta, 2^\Theta)$ is a loss function and τ is a (randomized) decision function, i.e. an estimator. Since we do not want to discriminate between different elements θ_1 and θ_2 of one H, it is natural to choose a loss function which does only depend on H and not on the specific θ; that is, we have a loss function

$$(W_H)_{H\in\mathcal{H}} \;\subset\; \mathcal{L}_\infty(\mathcal{H}, 2^{\mathcal{H}})$$

Furthermore, the decision space changes from Θ to $\mathcal{H}$ and the risk function becomes

$$\mathcal{H} \;\to\; \mathbb{R}\,, \qquad H \;\mapsto\; \sup_{\theta \in H} \sup_{P'_\theta \in \mathcal{M}'_\theta} \int_{\mathcal{X}} \int_{\mathcal{H}} W_H(\hat{H})\, \tau_x(d\hat{H})\, P'_\theta(dx) \tag{6.5}$$

for an estimator τ. Next, put

$$\mathcal{M}'_H \;:=\; \mathit{c\ell co} \bigcup_{\theta\in H} \mathcal{M}'_\theta\,, \qquad \forall\, H \in \mathcal{H}$$

where $\mathit{c\ell co}$ denotes the convex $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$-closure. That is, $\mathcal{M}'_H$ is the credal set of the coherent upper prevision $\overline{P}'_H$ defined by

$$\overline{P}'_H \;:\; \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}') \;\to\; \mathbb{R}\,, \qquad f \;\mapsto\; \sup_{\theta\in H} \overline{P}'_\theta[f]$$

According to Lemma 8.29, the risk function defined by (6.5) is equal to

$$\mathcal{H} \;\to\; \mathbb{R}\,, \qquad H \;\mapsto\; \sup_{P'_H \in \mathcal{M}'_H} \int_{\mathcal{X}} \int_{\mathcal{H}} W_H(\hat{H})\, \tau_x(d\hat{H})\, P'_\theta(dx) \tag{6.6}$$

and this function exactly coincides with the usual risk function defined in Section 3.2 if $(\overline{P}'_H)_{H\in\mathcal{H}}$ is our imprecise model. That is, discretizing Θ naturally leads to the imprecise model $(\overline{P}'_H)_{H\in\mathcal{H}}$, where $\mathcal{H}$ is a finite index set.

Of course, a thoughtless application of this discretization may lead to very bad results. This is because discretizing Θ means that we do not want to discriminate between different elements θ_1 and θ_2 of one H and, therefore, it is crucial to choose a sensible partition of Θ in order to get sensible results – the more since choosing a partition of Θ means choosing the statistical purpose.

So far, this method can be justified well. However, problems arise in applications since it is a necessary assumption for the applications presented in the present book that credal sets are given by a finite number of restrictions; cf. e.g. (5.30). However, even if there is a finite set $\mathcal{K} \subset \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ such that

$$\mathcal{M}'_\theta \;=\; \big\{ P'_\theta \in \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}') \;\big|\; P'_\theta[f] \le \overline{P}'_\theta[f] \quad \forall\, f \in \mathcal{K} \big\}$$

it does not seem to be clear if assumption (5.30) is fulfilled for $\mathcal{M}'_H$ which would be necessary to successfully work with $\mathcal{M}'_H$ in our applications. An ad hoc solution of this problem is to use the credal set

$$\hat{\mathcal{M}}'_H \;=\; \big\{P'_H \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \;\big|\; P'_H[f] \le \overline{P}'_H[f] \quad \forall\, f \in \mathcal{K}\big\}$$

as an "approximation" of $\mathcal{M}'_\theta$. It is easy to see that

$$\mathcal{M}'_\theta \;\subset\; \hat{\mathcal{M}}'_H$$

After that, $(\mathcal{X},\mathcal{A}')$ may be discretized according to Subsection 5.4.2 where the index set is given by $\mathcal{H}$.

6.3 A minimum distance estimator for imprecise models

In short, we are faced with a random sample

$$x_1\,,\;\ldots\;,\;x_n$$

from a precise distribution P'_θ which is unknown. It is only known that P'_θ is contained in a credal set $\mathcal{M}'_\theta$. The parameter θ is also unknown and should be estimated.
The idea of the presented minimum distance estimator is very simple:

> The data $x_1, \ldots, x_n$ are used to build the empirical measure
>
> $$\mathbb{P}^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$
>
> Then, the minimum distance estimator is that $\hat\theta \in \Theta$ such that $\mathbb{P}^{(n)}$ lies next to $\mathcal{M}'_{\hat\theta}$. That is, we calculate the distance between $\mathbb{P}^{(n)}$ and $\mathcal{M}'_\theta$ for every $\theta \in \Theta$ and pick that $\hat\theta$ where the distance is minimal.

This estimator will not be optimal in the general decision theoretic setup and the present section fails to proof any optimality result – the present section even does not make any attempt to derive such an optimality result. Admittedly, this is criticizable since, as a rule, every promoted statistical procedure should be justified by an appropriate optimality criterion.

On the other hand, even small numbers of observations (e.g. $n = 10$) usually lead to models which are so extensive that calculating optimal estimators is excluded because of exceedingly high computational efforts – at least as measured by the present state of research. So, the best that we can hope for at the moment are optimal estimators which cannot be calculated or estimators which can be calculated and behave reasonably well. The purpose of the present section is to develop such an estimator which can be calculated in real applications. The proposed minimum distance estimator fulfills this practical need in many situations. Furthermore, the asymptotic results of Section 6.4 confirm that the estimator behaves reasonably well in terms of asymptotic statistics, and the simulation study in Section 6.6 demonstrates its applicability.

In order to define the estimator in a mathematical rigorous way, the setup developed in Section 6.2 is used:

Let Ω be a set with σ-algebra $\mathcal{F}$ and $\mathcal{X}$ be a set with σ-algebra[9] $\mathcal{A}'$. Let Θ be any index set. There is no need to assume finiteness of Θ at the moment – such an assumption will only be used for concrete computations in Section 6.5.
Let $(\overline{U}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\Omega,\mathcal{F})$ with corresponding family of credal sets $(\mathcal{U}_\theta)_{\theta\in\Theta}$. The observations $x_1, \ldots, x_n$ are modeled via random variables

$$X_i \;:\; \Omega \;\to\; \mathcal{X}\;, \qquad\quad i\in\{1,\ldots,n\}$$

It is assumed that $X_1, \ldots, X_n$ are independent uniformly distributed with respect to an unknown probability charge $U_\theta \in \mathcal{U}_\theta$.
Therefore, we have an imprecise model $(P'_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A}')$ with corresponding credal sets

$$\mathcal{M}'_\theta \;=\; \big\{X(U_\theta)\;\big|\;U_\theta\in\mathcal{U}_\theta\big\}\,, \qquad \theta\in\Theta\,;$$

and the random variables

$$X_1\,,\;\ldots\,,\;X_n \quad \sim_{\text{i.i.d.}} \quad P'_\theta$$

are independent identically distributed according some precise distribution P'_θ which may be any element of the credal set of $\overline{P}'_\theta$. The task is to estimate the unknown parameter $\theta\in\Theta$.

The following fundamental assumption is made:

Assumption 6.3 *There is a finite subset* $\mathcal{K}=\{f_1,\ldots,f_s\}\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ *such that*

$$\mathcal{M}'_\theta \;=\; \big\{P'_\theta\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\;\big|\;P'_\theta[f]\le\overline{P}'_\theta[f]\quad\forall\,f\in\mathcal{K}\big\} \tag{6.7}$$

for every $\theta\in\Theta$. *Furthermore, it is assumed that*

$$\overline{P}'_\theta[f]-\underline{P}'_\theta[f]\;>\;0 \qquad\quad \forall\,f\in\mathcal{K} \tag{6.8}$$

where $\underline{P}'_\theta$ *is the corresponding lower coherent prevision.*[10]

Such assumptions have also been made in Subsection 5.4.2. As has already been stated there, these assumptions can be justified as follows: Practitioners will very often only be able to specify concrete upper previsions for a finite number of functions and this directly leads to models satisfying Assumption (6.7). In particular, this will often be true for expert systems. There, it is a natural proceeding to ask some experts about their prevision (or expectation) on some specific events, experiments, gambles, assets etc. – and this can only be done for a finite number of such objects.
Furthermore, Section 5.2 tell us that using models of form (6.7) which violate (6.8) is dangerous because these models are potentially most instable. Therefore, those models which violate (6.8) generally should be avoided anyway.

[9] In order to derive asymptotic results later on, some parts of the investigations are concerned with σ-additive probability measures and, therefore, we have to consider σ-algebras. This does not provide difficulties because an imprecise model on an algebra can always be extended to an imprecise model on a σ-algebra by means of a natural extension.

[10] That is, $\underline{P}'_\theta[f'] \;=\; \inf\limits_{P'_\theta\in\mathcal{M}'_\theta} P'_\theta[f']$

Note that these assumptions rule out classical probability measures. One of the main goals K. Weichselberger had when he developed his theory of imprecise probabilities (F-probabilities) was: "As a special case, classical probability must fit into this theory." [11] This means that – as a fundamental property – every probability measure is also an F-probability (and a coherent upper prevision). However, F-probabilities and coherent upper previsions which fulfill the above assumptions cannot coincide with probability measures. Accordingly, the following investigations do not apply to classical probability theory as a special case. That is, we deal with a *strictly* imprecise setup. As will be seen, this turns out to be an advantage here because the minimum distance estimator is based on the total variation distance. While working with total variation distances provides some difficulties in classical probability theory these difficulties cannot occur in our strictly imprecise setup; cf. Section 6.4.

Now, it is possible to define the *empirical measure* in this setup. The empirical measure $\mathbb{P}^{(n)}$ is the map

$$\mathbb{P}^{(n)} \;:\; \Omega \;\to\; \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\,, \qquad \omega \;\mapsto\; \mathbb{P}^{(n)}_\omega \;=\; \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}$$

where δ_{x_i} denotes the Dirac measure in $x_i \in \mathcal{X}$. Note that

$$\mathbb{P}^{(n)}[f'] \;:\; \Omega \;\to\; \mathbb{R}\,, \qquad \omega \;\mapsto\; \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}[f'] \;=\; \frac{1}{n}\sum_{i=1}^n f'\big(X_i(\omega)\big)$$

is a (bounded) random variable for every $f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ and

$$\Omega \times \mathcal{A}' \;\to\; \mathbb{R}\,, \qquad (\omega, A') \;\mapsto\; \mathbb{P}^{(n)}_\omega[I_{A'}]$$

is a Markov kernel.
The following notation will also be used: For every $x = (x_1,\dots,x_n) \in \mathcal{X}^n$, the probability measure on $(\mathcal{X},\mathcal{A}')$ defined by

$$\mathbb{P}^{(n)}_x[f'] \;:=\; \frac{1}{n}\sum_{i=1}^n f'\big(x_i\big) \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

is denoted by $\mathbb{P}^{(n)}_x$.

In order to define a minimum distance estimator, we have to choose a suitable notion of "distance" between a measure P_0' and a coherent upper prevision $\overline{P}'$ on $(\mathcal{X},\mathcal{A}')$ now.
Appropriately to the sensitivity analyst's point of view, the distance will be defined as

$$\inf_{P' \in \mathcal{M}'} d(P_0', P')$$

where d is a suitable metric on $\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')$.
Since bounded charges $\mu' \in \mathrm{ba}(\mathcal{X},\mathcal{A}')$ are mainly regarded as bounded linear operators on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ within the theory of imprecise probabilities, it seems to be most natural to choose the operator norm for d; that is,

$$d(P_0', P') \;=\; \|P_0' - P'\| \;=\; \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\big|\, P_0'[f'] - P'[f']\,\big|}{\|f'\|}$$

[11] (Weichselberger, 2000, p. 149f)

and we put

$$\left\|P'_0 - \overline{P}'\right\| \;:=\; \inf_{P' \in \mathcal{M}'} \|P'_0 - P'\| \tag{6.9}$$

Though this is not a norm (because of the different roles of P'_0 and $\overline{P}'$) this notation is sensible. Particularly, in the special case that $\overline{P}'$ is a precise prevision (i.e. a probability charge), the definition in (6.9) reduces to the usual operator norm in $\mathrm{ba}(\mathcal{X}, \mathcal{A}')$.

Next, the minimum distance estimator can be defined: The *minimum distance estimator* $\hat{\theta}'_n$ is

$$\hat{\theta}'_n \;:\; \mathcal{X}^n \;\to\; \Theta\,, \qquad x \;\mapsto\; \arg\min_{\theta \in \Theta} \left\|\mathbb{P}^{(n)}_x - \overline{P}'_\theta\right\|$$

Recall from Section 2.3 that the operator norm in $\mathrm{ba}(\mathcal{X}, \mathcal{A}')$ is equal to the total variation. Therefore, the minimum distance estimator is based on the total variation norm. As shown in the following section, the annoying properties of the total variation norm with respect to the empirical measure in classical statistics completely disappear in the above developed setup based on imprecise probabilities.

6.4 Asymptotic properties of the estimator

The setup and the notations of Section 6.3 are still valid. The present Section provides some theoretical justification of the minimum distance estimator defined in Section 6.3 . This justification solely relies on asymptotic arguments. In order to apply such arguments, σ-additivity becomes important. Therefore, it would be desirable that the coherent upper previsions were even upper expectations because, for upper expectations, considering σ-additive probability measures is sufficient in most situations; confer Section 2.4. Fortunately, coherent upper previsions which fulfill (6.7) and (6.8) are always upper expectations (cf. Section 2.4). This is the content of the following proposition:

Proposition 6.4 *Let $\underline{P}'_\theta$ be the lower coherent prevision which corresponds to $\overline{P}'_\theta$. Assume (6.7) and (6.8).*
Then, $\overline{P}'_\theta$ is an upper expectation, i.e.

$$\mathcal{M}'_\theta \cap \mathrm{ca}^+_1(\mathcal{X}, \mathcal{A}')$$

is dense in $\mathcal{M}'_\theta$ with respect to the $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ - topology.

Proof: Fix any $P'_0 \in \mathcal{M}'_\theta$, any $f'_0 \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$ and any $\varepsilon > 0$. Then, it is enough to show that there is a $P'_c \in \mathcal{M}'_\theta \cap \mathrm{ca}^+_1(\mathcal{X}, \mathcal{A}')$ such that

$$P'_c[f'_0] \;>\; P'_0[f'_0] - \varepsilon \tag{6.10}$$

because this implies

$$\sup\left\{P'_\theta[f'_0] \;\middle|\; P'_\theta \in \mathcal{M}'_\theta \cap \mathrm{ca}^+_1(\mathcal{X}, \mathcal{A}')\right\} \;=\; \overline{P}'_\theta[f'_0]$$

See also Proposition 2.21.

Put

$$\mathcal{K} \;=\; \{f_1, \dots, f_s\}$$

Assumption (6.8) implies that, for every $i \in \{1,\dots,s\}$, there is some $P_i' \in \mathcal{M}_\theta'$ such that

$$P_i'[f_i'] \;<\; \overline{P}_\theta'[f_i'] \tag{6.11}$$

Put

$$P_\alpha' \;=\; (1-\alpha)P_0' + \frac{\alpha}{s}\sum_{i=1}^{s} P_i'$$

for $\alpha \;=\; \dfrac{\varepsilon}{4\|f_0'\|} \wedge 1 \;\in\; (0,1]$.

Of course, convexity of $\mathcal{M}_\theta'$ implies $P_\alpha' \in \mathcal{M}_\theta'$ but, even more,

$$\begin{aligned} P_\alpha'[f_j] \quad &= \quad (1-\alpha)P_0'[f_j] + \frac{\alpha}{s}\sum_{i=1}^{s} P_i'[f_j] \;\leq \\ &\leq \quad (1-\alpha)\overline{P}_\theta'[f_j] + \frac{\alpha}{s}\Big(P_j'[f_j] + (s-1)\overline{P}_\theta'[f_j]\Big) \;< \\ &\overset{(6.11)}{<} \quad (1-\alpha)\overline{P}_\theta'[f_j] + \frac{\alpha}{s}\Big(\overline{P}_\theta'[f_j] + (s-1)\overline{P}_\theta'[f_j]\Big) \;=\; \overline{P}_\theta'[f_j] \end{aligned}$$

Hence,

$$\varepsilon_0 \;:=\; \min\Big\{\,\frac{\varepsilon}{2}\,,\; \overline{P}_\theta'[f_1] - P_\alpha'[f_1]\,,\;\dots\;,\; \overline{P}_\theta'[f_s] - P_\alpha'[f_s]\Big\} \;>\; 0$$

Furthermore,

$$\begin{aligned} \big|P_\alpha'[f_0'] - P_0'[f_0']\big| \;&\leq\; \alpha\big|P_0'[f_0']\big| + \frac{\alpha}{s}\sum_{j=1}^{s}\big|P_j'[f_0']\big| \;\leq \\ &\leq\; \alpha\|f_0'\| \;+\; \frac{\alpha}{s}\sum_{j=1}^{s}\|f_0'\| \;=\; 2\alpha\|f_0'\| \;\leq\; \frac{\varepsilon}{2} \end{aligned} \tag{6.12}$$

For every $j \in \{1,\dots,s\}$, put

$$\Lambda_j \;:\; \mathrm{ba}(\mathcal{X},\mathcal{A}') \;\to\; \mathbb{R}\,, \qquad \mu' \;\mapsto\; \mu'[f_j]$$

and put

$$\Lambda_0 \;:\; \mathrm{ba}(\mathcal{X},\mathcal{A}') \;\to\; \mathbb{R}\,, \qquad \mu' \;\mapsto\; \mu'[f_0']$$

Since these maps are $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ - continuous,

$$B_0 \;:=\; \bigcap_{j=1}^{s} \Lambda_j^{-1}\Big(\big(-\infty\,,\,P_\alpha'[f_j]+\varepsilon_0\big)\Big) \;\cap\; \Lambda_0^{-1}\Big(\big(P_\alpha'[f_0']-\varepsilon_0\,,\,\infty\big)\Big)$$

is an $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ - open neighborhood of P_α' and, therefore, Theorem 2.11 b) implies the existence of some

$$P_c' \;\in\; B_0 \cap \mathrm{ca}_1^+(\mathcal{X},\mathcal{A}')$$

Hence, it follows from

$$\begin{aligned} P_c'[f_j] \;&\leq\; P_\alpha'[f_j] + \varepsilon_0 \;\leq\; P_\alpha'[f_j] + \big(\overline{P}_\theta'[f_j] - P_\alpha'[f_j]\big) \;= \\ &=\; \overline{P}_\theta'[f_j] \qquad\quad \forall\, j \in \{1,\dots,s\} \end{aligned}$$

that $P_c' \;\in\; \mathcal{M}_\theta' \cap \mathrm{ca}_1^+(\mathcal{X},\mathcal{A}')$. Furthermore,

$$P_c'[f_0'] \;>\; P_\alpha'[f_0'] - \varepsilon_0 \;\overset{(6.12)}{\geq}\; P_0'[f_0'] - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} \;=\; P_0'[f_0'] - \varepsilon$$

□

The following example shows that assumption (6.8) cannot be omitted:

Example 6.5 *Put* $\mathcal{X} = (0,1]$ *and let* $\mathcal{A}'$ *be the Borel-*σ*-algebra on* $(0,1]$. *Let* $\mathcal{K} = \{f_1\}$ *where*

$$f_1 \;:\; (0,1] \;\to\; \mathbb{R}\,, \qquad x \;\mapsto\; f_1(x) \;=\; x$$

and take

$$\overline{P}'_\theta[f_1] \;=\; 0$$

It is not obvious that this is defined well because this is equivalent with the existence of a probability charge $P'_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')$ *such that*

$$P'_\theta[f_1] \;=\; 0 \tag{6.13}$$

and (6.13) is very strange – at least for readers who are used to σ*-additive probability measure – because* f_1 *is strictly positive! In particular, note that (6.13) implies*

$$P'_\theta\big((\varepsilon,1]\big) \;=\; 0 \qquad \forall\,\varepsilon > 0 \qquad\quad \textit{but} \qquad\quad P'_\theta\big((0,1]\big) \;=\; 1$$

So, where does P'_θ *put its mass on?*

It is shown now, that such a probability charge P'_θ *really does exist. Obviously,* P'_θ *cannot be* σ*-additive. According to (Hoffmann-Jørgensen, 1994a, p. xxxvii), the existence of a probability charge on a* σ*-algebra which is not a* σ*-additive probability measure is equivalent to a certain form of the axiom of choice and, therefore, it is clear from the first that the following proof will need this form of the axiom of choice.*

It is easy to see that

$$T \;:\; \big\{a\!\cdot\! f_1 \;\big|\; a \in \mathbb{R}\big\} \;\to\; \mathbb{R}\,, \qquad f' \;\mapsto\; T(f') \;=\; 0$$

is a (norm-)continuous linear operator such that

$$T(f') \;\le\; \sup f' \qquad\quad \forall\, f' \in \big\{a\!\cdot\! f_1 \;\big|\; a \in \mathbb{R}\big\}$$

Therefore, it follows from the Hahn-Banach Theorem (Dunford and Schwartz, 1958, Theorem II.3.10) that T *may be extended to a continuous linear operator on* $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ *such that*

$$T(f') \;\le\; \sup f' \qquad\quad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

and Theorem 2.4 implies the existence of a bounded charge $\mu' \in \mathrm{ba}(\mathcal{X},\mathcal{A}')$ *such that*

$$\mu'[f'] \;\;=\;\; T(f')] \qquad\quad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

Especially, $\mu'[f_1] \;=\; 0\,$. *It only remains to show that* μ *is even a probability charge. This follows from*

$$\mu(A') \;\;=\;\; -T(-I_{A'}] \;\ge\; -\sup\big(-I_{A'}\big) \;\ge\; 0 \qquad \forall\, A' \in \mathcal{A}'$$

and

$$1 \;=\; -\sup\big(-I_{\mathcal{X}}\big) \;\le\; -T(-I_{\mathcal{X}}) \;=\; \mu(\mathcal{X}) \;=\; T(I_{\mathcal{X}}) \;\le\; \sup I_{\mathcal{X}} \;=\; 1$$

That is, $P'_\theta = \mu$ *is a probability charge which fulfills (6.13).*

Though the above mentioned certain form of the axiom of choice is not visible in the proof, it has nevertheless been used – it is associated with the Hahn-Banach Theorem.

As already mentioned in Section 6.3, the use of the total variation norm together with the empirical measure is not unproblematic in classical statistics. For example, consider a random sample from a standard normal distribution

$$X_1\,,\;\dots\;,\;X_n \quad \sim_{\text{i.i.d.}} \quad \mathcal{N}(0,1) \;:=\; P$$

Recall the notations

$$\mathbb{P}^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \qquad \text{and} \qquad \mathbb{P}^{(n)}_{\omega} \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i(\omega)} \qquad \forall\,\omega\in\Omega$$

for the empirical measure. Then, several distances d provides the desirable property that

$$d\big(\mathbb{P}^{(n)}_{\omega},P\big) \quad \xrightarrow[n\to\infty]{} \quad 0 \qquad P(d\omega)\;-\;\text{a.s.} \tag{6.14}$$

This is e.g. true for the Kolmogorov-Smirnov distance and the Cramér-von Mises distance; however, it is not true for the total variation norm. In order to see this, fix any $\omega\in\Omega$ and put

$$f\;:\;\mathcal{X}\;\to\;\mathbb{R}\,, \qquad x\;\mapsto\; I_{\{X_1(\omega),\dots,X_n(\omega)\}}(x) - I_{\mathcal{X}\setminus\{X_1(\omega),\dots,X_n(\omega)\}}(x)$$

Then, we have $\mathbb{P}^{(n)}_{\omega}[f]=1$ and $P[f]=-1$ and, therefore,

$$\|\,\mathbb{P}^{(n)}_{\omega}-P\,\| \;=\; 2 \qquad \forall\,n\in\mathbb{N} \qquad \forall\,\omega\in\Omega$$

which is the worst possible violation of (6.14). However, Theorem 6.6 below states that this annoying difficulty totally disappears in the imprecise probability setup summarized in Section 6.3. If we replace P by a coherent upper prevision $\overline{P}'$ satisfying assumptions (6.7) and (6.8), we get

$$\big\|\mathbb{P}^{(n)}-\overline{P}'\big\| \quad \xrightarrow[n\to\infty]{} \quad 0 \qquad P_0'\;-\;\text{a.s.}^* \tag{6.15}$$

for every probability measure P_0' in the credal set of $\overline{P}'$. In (6.15), writing a.s.* instead of a.s. indicates that there may be some problems concerning measurability because, in general, the map

$$\omega \quad\mapsto\quad \big\|\mathbb{P}^{(n)}_{\omega}-\overline{P}'\big\| \;=\; \inf_{P'\in\mathcal{M}'}\;\sup_{f'\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')}\;\frac{\big|\,\mathbb{P}^{(n)}_{\omega}[f']-P'[f']\,\big|}{\|f'\|}$$

is not measurable.

In order to stay mathematically rigorously, consider the following notations which are in accordance with the setup in Section 6.3:

Let U_0 be a probability measure on $(\Omega,\mathcal{F})$ and let

$$X_i\;:\;\Omega\;\to\;\mathcal{X}\,, \qquad i\in\{1,\dots,n\}$$

be random variables which are independent identically distributed with respect to U_0 . It is assumed that the image

$$P_0' \;:=\; X_i(U_0)$$

is any element of the credal set $\mathcal{M}'$ which belongs to a coherent upper prevision $\overline{P}'$ on $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$. Of course, P_0 does not depend on i then. Again, we assume the validity of (6.7) and (6.8). That is, the credal set $\mathcal{M}'$ is given by

$$\mathcal{M}' \;=\; \left\{P' \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}') \,\middle|\, P'[f] \le \overline{P}'[f] \quad \forall\, f \in \mathcal{K}\right\} \tag{6.16}$$

where $\mathcal{K} = \{f_1, \ldots, f_s\}$ is assumed to be a finite subset of $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ and

$$\overline{P}'[f] - \underline{P}'[f] \;>\; 0 \qquad \forall\, f \in \mathcal{K} \tag{6.17}$$

More precisely, (6.15) means

$$\big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \;\xrightarrow[n\to\infty]{}\; 0 \qquad U_0(d\omega) \;-\; \text{a.s.}^* \tag{6.18}$$

where $\omega \mapsto \big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\|$ denotes the map

$$\Omega \;\to\; \mathbb{R}\,, \qquad \omega \;\mapsto\; \inf_{P' \in \mathcal{M}'} \; \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\left|\frac{1}{n}\sum_{i=1}^n f'\big(X_i(\omega)\big) \;-\; P'[f']\right|}{\|f'\|}$$

Taking the measurability issues indicated by the asterisk in a.e.* into account, (6.18) precisely means:

> There is a sequence of $\mathcal{A}'/\mathbb{R}$- measurable random variables
>
> $$\Delta_n \;:\; \Omega \;\to\; \mathbb{R}\,, \qquad \omega \;\mapsto\; \Delta_n(\omega)$$
>
> such that
>
> $$\big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \;\le\; \Delta_n(\omega) \qquad \forall\, \omega \in \Omega \qquad\quad \forall\, n \in \mathbb{N}$$
>
> and
>
> $$\Delta_n \;\xrightarrow[n\to\infty]{}\; 0 \qquad U_0 \;-\; \text{a.s.}$$

Confer e.g. (van der Vaart, 1998, §18) or (van der Vaart and Wellner, 1996, §1.9) for this definition of almost sure convergence of unmeasurable maps.
Now, the already pronounced theorem can be formulated:

Theorem 6.6 *In the setup of the present section, assume that*

$$U_0 \;\in\; \mathrm{ca}_1^+(\Omega,\mathcal{F}) \tag{6.19}$$

Let $\overline{P}'$ be a coherent upper prevision with credal set $\mathcal{M}'$ such that

$$P_0' \;=\; X_i(U_0) \;\in\; \mathcal{M}'$$

Assume that $\mathcal{M}'$ fulfills (6.16) and (6.17).
Then,

$$\big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \;\xrightarrow[n\to\infty]{}\; 0 \qquad U_0(d\omega) \;-\; a.s.^*$$

The proof of Theorem 6.6 needs some preparations which are even interesting from its own: Lemma 6.7 provides a different description of the distance

$$\|\mathbb{P}_\omega^{(n)} - \overline{P}'\|$$

and Proposition 6.8 provides a bound on this distance which does only depend on the values

$$\mathbb{P}_\omega^{(n)}[f_i]\,,\ \ \overline{P}'[f_i] \quad\text{and}\quad \underline{P}'[f_i] \qquad\quad \text{for} \quad i \in \{1,\dots,s\}$$

This is a nice property for practical applications because these values are already known or can easily be calculated. In particular, Proposition 6.8 is important for the practical implementation of the minimum distance estimator because it will follow from Proposition 6.8 that the proposed algorithm is correct.

Lemma 6.7 *Let $\overline{Q}$ be a coherent upper prevision on $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ with corresponding credal set $\mathcal{N}$ on $(\mathcal{Y},\mathcal{B})$ and let Q_0 be a propability charge on $(\mathcal{Y},\mathcal{B})$.*
Let G be a subset of $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ such that

- *G is convex*
- *$g \in G \quad\Rightarrow\quad -g \in G$*
- *G is bounded: $\sup_{g\in G}\|g\| < \infty$*

Then,

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}\big|Q_0[g] - Q[g]\big| \;=\; \sup_{g\in G}\; Q_0[g] - \overline{Q}[g] \tag{6.20}$$

In particular,

$$\big\|Q_0 - \overline{Q}\big\| \;=\; \sup_{g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})} \frac{Q_0[g] - \overline{Q}[g]}{\|g\|}$$

That is, the distance $\big\|Q_0 - \overline{Q}\big\|$ exactly coincides with the operator norm if we consider

$$\mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \;\to\; \mathbb{R}\,, \qquad g \;\mapsto\; Q_0[g] - \overline{Q}[g]$$

as a (non-linear) operator.

Proof of Lemma 6.7: Equation (6.20) obviously coincides with

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}\big|Q_0[g] - Q[g]\big| \;=\; \sup_{g\in G}\inf_{Q\in\mathcal{N}}\; Q_0[g] - Q[g] \tag{6.21}$$

In (6.21) the inequality “$\geq$” is trivial and, therefore, it only remains to proof the inequality “$\leq$” in (6.21).

In order to prove this, firstly, fix any $Q \in \mathcal{N}$ and any $g_0 \in G$. In case of $Q_0[g_0] \geq Q[g_0]$, we have

$$\big|Q_0[g_0] - Q[g_0]\big| \;=\; Q_0[g_0] - Q[g_0] \;\leq\; \sup_{g\in G} Q_0[g] - Q[g]$$

and in case of $Q_0[g_0] \leq Q[g_0]$, we also have

$$\big|Q_0[g_0] - Q[g_0]\big| \;=\; Q_0[-g_0] - Q[-g_0] \;\leq\; \sup_{g\in G} Q_0[g] - Q[g]$$

since $-g_0 \in G$. Hence, it follows that

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}\big|Q_0[g]-Q[g]\big| \;=\; \inf_{Q\in\mathcal{N}}\sup_{g\in G}\;Q_0[g]-Q[g] \tag{6.22}$$

In order to show that inf and sup may be interchanged in (6.22), a minimax theorem is applied for

$$\Gamma\;:\;\mathcal{N}\times G\;\to\;\mathbb{R}\,,\qquad (Q,g)\;\mapsto\;Q_0[g]-Q[g]$$

To this end, note that $\mathcal{N}$ is $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$-compact and, for every $g\in G$, $Q\mapsto\Gamma(Q,g)$ is convex and $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$-continuous. In addition, $g\mapsto\Gamma(Q,g)$ is concave for every $Q\in\mathcal{N}$. So, it follows from the minimax theorem (Fan, 1953, Theorem 2) that

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}\;\Gamma(Q,g)\;=\;\sup_{g\in G}\inf_{Q\in\mathcal{N}}\;\Gamma(Q,g)$$

Together with (6.22), this implies (6.21) and (6.20).

In particular, for

$$G\;=\;\big\{g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})\;\big|\;\|g\|\le 1\big\}$$

it follows that

$$\begin{aligned}\big\|Q_0-\overline{Q}\big\| \;&=\; \inf_{Q\in\mathcal{N}}\sup_{g\in G}\big|Q_0[g]-Q[g]\big| \;\overset{(6.20)}{=}\; \sup_{g\in G}\;Q_0[g]-\overline{Q}[g]\;=\\ &=\;\sup_{g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})}\;\frac{Q_0[g]-\overline{Q}[g]}{\|g\|}\end{aligned}$$

□

Proposition 6.8 *Let $\overline{Q}$ be a coherent upper prevision on $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ with credal set*

$$\mathcal{N}\;=\;\big\{Q\in\mathrm{ba}_1^+(\mathcal{Y},\mathcal{B})\;\big|\;Q[g]\le\overline{Q}[g]\quad\forall\,g\in\mathcal{G}\big\} \tag{6.23}$$

where $\mathcal{G}=\{g_1,\dots,g_s\}$ is a finite subset of $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$. Assume that

$$\overline{Q}[g_i]-\underline{Q}[g_i]\;>\;0\qquad\forall\,i\in\{1,\dots,s\} \tag{6.24}$$

Then, for every probability charge $Q_0\in\mathrm{ba}_1^+(\mathcal{Y},\mathcal{B})$,

$$\|Q_0-\overline{Q}\|\;\;\le\;\;2\cdot\sum_{i=1}^{s}\frac{\big(Q_0[g_i]-\overline{Q}[g_i]\big)^+}{\overline{Q}[g_i]-\underline{Q}[g_i]} \tag{6.25}$$

Proof of Proposition 6.8 If there is any $i\in\{1,\dots,s\}$ such that $Q_0[g_i]-\overline{Q}[g_i]\ge\overline{Q}[g_i]-\underline{Q}[g_i]$ then (6.25) is trivially fulfilled and nothing remains to be proven.

Therefore, it can be assumed that

$$Q_0[g_i]-\overline{Q}[g_i]\;<\;\overline{Q}[g_i]-\underline{Q}[g_i]\qquad\forall\,i\in\{1,\dots,s\} \tag{6.26}$$

Without loss of generality, we may assume that the elements of $\mathcal{G}$ are indexed in such a way that there is a $r\in\{0,\dots,s\}$ such that

$$Q_0[g_i]\;>\;\overline{Q}[g_i]\quad\;\forall\,i\le r\qquad\text{and}\qquad Q_0[g_i]\;\le\;\overline{Q}[g_i]\quad\;\forall\,i>r$$

Putting

$$\varepsilon_i \;:=\; \frac{\big(Q_0[g_i]-\overline{Q}[g_i]\big)^+}{\overline{Q}[g_i]-\underline{Q}[g_i]} \qquad \forall\, i\le r\,, \tag{6.27}$$

(6.26) implies $0<\varepsilon_i<1$ for every $i\in\{1,\dots,r\}$.

Let $\overline{Q}_0$ be the coherent upper prevision with credal set

$$\mathcal{N}_0 \;=\; \Big\{Q\in \mathrm{ba}_1^+(\mathcal{Y},\mathcal{B})\;\Big|\; Q[g]\;\le\;\max\big\{\overline{Q}[g]\,,\,Q_0[g]\big\}\quad\forall\, g\in\mathcal{G}\Big\}$$

Then, it follows from $Q_0\in\mathcal{N}_0$ and $\mathcal{N}\subset\mathcal{N}_0$ that

$$\overline{Q}_0[g]\;=\;\max\big\{\overline{Q}[g]\,,\,Q_0[g]\big\}\qquad \forall\, g\in\mathcal{G}$$

and, together with (6.27), this implies

$$\overline{Q}[g_i]\;\le\;\overline{Q}_0[g_i]\;=\;\overline{Q}[g_i]+\varepsilon_i\big(\overline{Q}[g_i]-\underline{Q}[g_i]\big)\qquad\forall\, i\le r$$

and

$$\overline{Q}[g_i]\;=\;\overline{Q}_0[g_i]\qquad\forall\, i>r$$

Since $\overline{Q}$ and $\overline{Q}_0$ may be considered as natural extensions of coherent upper previsions on $\mathcal{G}$, Proposition 5.1 is applicable and yields

$$\overline{Q}[g]\;\le\;\overline{Q}_0[g]\;\le\;\overline{Q}[g]+\varepsilon\big(\sup g-\inf g\big)\;\le\;\overline{Q}[g]+2\varepsilon\|g\|\qquad\forall\, g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$$

for $\varepsilon\;=\;\varepsilon_1+\dots+\varepsilon_r\;>0$.

Put

$$G\;:=\;\big\{g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})\;\big|\;\|g\|\le 1\big\}$$

Then, Lemma 6.7 implies

$$\begin{aligned}\|Q_0-\overline{Q}\| \;&=\; \sup_{g\in G} Q_0[g]-\overline{Q}[g]\;\le\;\sup_{g\in G}\overline{Q}_0[g]-\overline{Q}[g]\;\le\\ &\le\;\sup_{g\in G}\overline{Q}[g]+2\varepsilon\|g\|-\overline{Q}[g]\;\le\;\sup_{g\in G}2\varepsilon\|g\|\;=\\ &=\;2\cdot\sum_{i=1}^{r}\varepsilon_i\;=\;2\cdot\sum_{i=1}^{s}\frac{\big(Q_0[g_i]-\overline{Q}[g_i]\big)^+}{\overline{Q}[g_i]-\underline{Q}[g_i]}\end{aligned}$$

□

Of course, (6.23) is, in general, a very bad bound. However, if Q_0 is equal to the empirical measure $\mathbb{P}^{(n)}$ and the true distribution lies in the credal set of $\overline{Q}=\overline{P}'$, then the law of large numbers yields

$$\big(Q_0[g_i]-\overline{Q}[g_i]\big)^+\;\xrightarrow[n\to\infty]{}\;0$$

Therefore, bound (6.23) provides valuable information for increasing numbers of observations. Since Theorem 6.6 is about the asymptotic behaviour of the distance $\|\mathbb{P}^{(n)}-\overline{P}'\|$, bound (6.23) serves as the cornerstone of the proof.

Proof of Theorem 6.6 For every $\omega \in \Omega$, an application of Proposition 6.8 yields

$$\big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \;\leq\; 2 \cdot \sum_{i=1}^{s} \frac{\big(\mathbb{P}_\omega^{(n)}[f_i] - \overline{P}'[f_i]\big)^+}{\overline{P}'[f_i] - \underline{P}'[f_i]} \tag{6.28}$$

The map

$$\Omega \;\to\; \mathbb{R}\,, \qquad \omega \;\mapsto\; 2 \cdot \sum_{i=1}^{s} \frac{\big(\mathbb{P}_\omega^{(n)}[f_i] - \overline{P}'[f_i]\big)^+}{\overline{P}'[f_i] - \underline{P}'[f_i]}$$

is measurable with respect to $\mathcal{F}$ and $\mathbb{B}$.

For every $i \in \{1, \dots, s\}$, the strong law of large numbers (Hoffmann-Jørgensen, 1994a, § 4.12) and the transformation theorem (Hoffmann-Jørgensen, 1994a, § 3.15) implies the existance of a U_0-set $N_i \in \mathcal{F}$ such that

$$\begin{aligned}\mathbb{P}_\omega^{(n)}[f_i] \;&=\; \frac{1}{n}\sum_{j=1}^{n} f_i \circ X_j(\omega) \;\xrightarrow[n\to\infty]{}\; \int_\Omega f_i \circ X_1(\omega)\, U_0(d\omega) \;=\\ &=\; \int_\Omega f_i(x)\, P_0'(dx) \;=\; P_0'[f_i] \;\leq\; \overline{P}'[f_i] \qquad \forall\, \omega \,\in\, \Omega \setminus N_i\end{aligned}$$

Therefore

$$N \;:=\; \bigcup_{i=1}^{s} N_i \;\in\; \mathcal{F}$$

is a U_0-set such that

$$2 \cdot \sum_{i=1}^{s} \frac{\big(\mathbb{P}_\omega^{(n)}[f_i] - \overline{P}'[f_i]\big)^+}{\overline{P}'[f_i] - \underline{P}'[f_i]} \quad \xrightarrow[n\to\infty]{} \quad 0 \qquad \forall\, \omega \,\in\, \Omega \setminus N$$

Together with (6.28), this proves Theorem 6.6. □

Theorem 6.6 states that the distance between the empirical measure and the coherent upper prevision converges to 0. However, the techniques developed for the proof can also be used to make some assertions about the rate of convergence:

Theorem 6.9 *Under the assumptions of Theorem 6.6, it follows that*

a) $$\big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \;=\; O\!\left(\sqrt{\frac{\ln\ln n}{n}}\right) \qquad U_0(d\omega) \,-\, a.s.^*$$

b) *In addition, assume that*

$$P_0'[f_i] \;<\; \overline{P}'_0[f_i] \qquad \forall\, i \in \{1, \dots, s\} \tag{6.29}$$

Then,

$$\lim_{n\to\infty} U_0^*\Big(\Big\{\omega \in \Omega \;\Big|\; \big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \,>\, 0\Big\}\Big) \;=\; 0$$

In part b), using the outer measure U_0^* instead of U_0 is due to the fact that the set

$$\Big\{\omega \in \Omega \;\Big|\; \big\|\mathbb{P}_\omega^{(n)} - \overline{P}'\big\| \,>\, 0\Big\}$$

will, in general, not be measurable.

Proof: For every $i \in \{1, \ldots, s\}$, put

$$h_i \;:=\; \frac{f_i}{\overline{P}'[f_i] - \underline{P}'[f_i]}$$

According to Proposition 6.8,

$$0 \;\le\; \|\mathbb{P}_\omega^{(n)} - \overline{P}'\| \;\le\; 2 \cdot \sum_{i=1}^{s} \big(\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}'[h_i]\big)^+ \qquad \forall\, \omega \in \Omega \tag{6.30}$$

a) Note that, for every $i \in \{1, \ldots, s\}$ and for every $\omega \in \Omega$,

$$\begin{aligned}\big(\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}'[h_i]\big)^+ \;&\le\; \big(\mathbb{P}_\omega^{(n)}[h_i] - P_0'[h_i]\big)^+ \;\le\; \big|\mathbb{P}_\omega^{(n)}[h_i] - P_0'[h_i]\big| \;=\\ &=\; \left|\frac{1}{n}\sum_{j=1}^{n}\Big(h_i \circ X_j(\omega) - P_0'[h_i]\Big)\right|\end{aligned}$$

Since

$$\int_\Omega h_i \circ X_j(\omega)\, U_0(d\omega) \;=\; P_0'[h_i]$$

the law of the iterated logarithm (Hoffmann-Jørgensen, 1994b, § 10.25) yields

$$\left|\frac{1}{n}\sum_{j=1}^{n}\Big(h_i \circ X_j(\omega) - P_0'[h_i]\Big)\right| \;=\; O\!\left(\sqrt{\frac{\ln\ln n}{n}}\right) \qquad U_0(d\omega) \,-\, \text{a.s.}^*$$

and, therefore,

$$\big(\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}'[h_i]\big)^+ \;=\; O\!\left(\sqrt{\frac{\ln\ln n}{n}}\right) \qquad U_0(d\omega) \,-\, \text{a.s.}^*$$

Together with (6.30), this implies the validity of part a).

b) For every $n \in \mathbb{N}$, put

$$A_n^{(1)} \;=\; \Big\{\omega \in \Omega \;\Big|\; \|\mathbb{P}_\omega^{(n)} - \overline{P}'\| \;>\; 0\Big\}$$

$$A_n^{(2)} \;=\; \Big\{\omega \in \Omega \;\Big|\; \sum_{i=1}^{s} \big(\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}'[h_i]\big)^+ \;>\; 0\Big\}$$

and, for $i \in \{1, \ldots, s\}$,

$$\begin{aligned}B_{i,n} \;&=\; \Big\{\omega \in \Omega \;\Big|\; \mathbb{P}_\omega^{(n)}[h_i] > \overline{P}'[h_i]\Big\} \;=\\ &=\; \Big\{\omega \in \Omega \;\Big|\; \mathbb{P}_\omega^{(n)}[h_i] - P_0'[h_i] \;>\; \overline{P}'[h_i] - P_0'[h_i]\Big\}\end{aligned}$$

Then, we have

$$A_n^{(1)} \;\overset{(6.30)}{\subset}\; A_n^{(2)} \;\subset\; \bigcup_{i=1}^{s} B_{i,n}$$

where $A_n^{(2)}$, $B_{1,n}$, ..., $B_{s,n} \in \mathcal{F}$. Finally,

$$U_0^*\big(A_n^{(1)}\big) \quad \leq \quad \sum_{i=1}^{s} U_0\big(B_{i,n}\big) \quad \xrightarrow[n\to\infty]{} 0$$

because

$$U_0\big(B_{i,n}\big) \quad \xrightarrow[n\to\infty]{} 0 \qquad \forall\, i \in \{1,\dots,s\}$$

follows from the strong law of large numbers (Hoffmann-Jørgensen, 1994a, § 4.12), assumption (6.29) and the fact that almost sure convergence implies convergence in probability; cf. (Hoffmann-Jørgensen, 1994a, § 3.25). □

Now, let us turn over to consistency of the minimum distance estimator

$$\hat{\theta}_n' \;:\; \mathcal{X}^n \;\to\; \Theta\,, \qquad x \;\mapsto\; \arg\min_{\theta\in\Theta} \big\|\mathbb{P}_x^{(n)} - \overline{P}_\theta'\big\|$$

In order to stay mathematically rigorous, it is more convenient to work on $(\Omega,\mathcal{F})$ than on the sample space $(\mathcal{X}^n, \mathcal{A}'^{\otimes n})$. Therefore, we most often consider the maps

$$\hat{\theta}_n \;:\; \Omega \;\to\; \Theta\,, \qquad \omega \;\mapsto\; \hat{\theta}_n'\big(X_1(\omega),\dots,X_n(\omega)\big)$$

instead of $\hat{\theta}_n'$. That is, we use the notation

$$\hat{\theta}_n \;=\; \hat{\theta}_n'\big(X_1,\dots,X_n\big) \;=\; \arg\min_{\theta\in\Theta} \big\|\mathbb{P}_\omega^{(n)} - \overline{P}_\theta'\big\|$$

Recall the setup presented in Section 6.3 now. That is, we have an imprecise model $(\overline{P}_\theta')_{\theta\in\theta}$ on the measurable space $(\mathcal{X},\mathcal{A}')$ with corresponding credal sets $(\mathcal{M}_\theta')_{\theta\in\Theta}$. With respect to a (fixed but totally unknown) probability measure U_0 on $(\Omega,\mathcal{F})$, the random variables

$$X_1\,,\;\dots\,,\;X_n \quad \sim_{\text{i.i.d.}} \quad P_{\theta_0}'$$

are independent identically distributed according some precise distribution

$$P_{\theta_0}' \;\in\; \mathcal{M}_{\theta_0}' \qquad \text{for some} \quad \theta_0 \in \Theta$$

The true $P_{\theta_0}' \in \mathcal{M}_{\theta_0}'$ is totally unknown and we only want to estimate θ_0. A true parameter θ_0 is any[12] $\theta_0 \in \Theta$ such that

$$P_{\theta_0}' \;\in\; \mathcal{M}_{\theta_0}'$$

In case of a finite set of parameters Θ, a sensible estimator $\hat{\theta}_n$ should – at least for large sizes of n – lead to small error probabilities

$$U_0^*\Big(\,P_0 \notin \mathcal{M}_{\hat{\theta}_n}'\Big)$$

so that

$$U_0^*\Big(\,P_0 \notin \mathcal{M}_{\hat{\theta}_n}'\Big) \quad \xrightarrow[n\to\,\infty]{} \quad 0$$

The minimum distance estimator fulfilles this requirement as stated in the following theorem. Again, using the outer measure U_0^* instead of U_0 is necessary because we do not assume

$$\Big\{\omega\in\Omega \;\Big|\; P_0 \notin \mathcal{M}_{\hat{\theta}_n(\omega)}'\Big\}$$

to be measurable.

[12] θ_0 is not assumed to be unique.

Theorem 6.10 *In the setup of the present section, assume that*

$$U_0 \;\in\; \mathrm{ca}_1^+(\Omega,\mathcal{F}) \tag{6.31}$$

Let $(\overline{P}'_\theta)_{\theta\in\theta}$ be an imprecise model on the measurable space $(\mathcal{X},\mathcal{A}')$ with corresponding family of credal sets $(\mathcal{M}'_\theta)_{\theta\in\Theta}$. The index set Θ is asumed to be finite. With respect to U_0 , the random variables

$$X_1\,,\;\ldots\;,\;X_n \quad \sim_{\text{i.i.d.}} \quad P'_{\theta_0}$$

are independent identically distributed according some precise distribution P'_{θ_0} such that

$$P'_{\theta_0} \;\in\; \mathcal{M}'_{\theta_0} \qquad \textit{for some} \quad \theta_0\in\Theta$$

Assume that $\mathcal{M}'_\theta$ fulfills (6.7) and (6.8) for every $\theta\in\Theta$.
Then,

$$U_0^*\Big(\,P_{\theta_0}\notin\mathcal{M}'_{\hat{\theta}_n}\Big) \quad \xrightarrow[n\to\,\infty]{} \quad 0 \tag{6.32}$$

Proof: Firstly, fix any $n\in\mathbb{N}$. For every $\theta\in\Theta$, put

$$A_\theta^{(n)} \quad := \quad \Big\{\omega\in\Omega\;\Big|\;\;\big\|\mathbb{P}_\omega^{(n)}-\overline{P}'_\theta\big\|\;\le\;\big\|\mathbb{P}_\omega^{(n)}-\overline{P}'_{\theta_0}\big\|\,\Big\}$$

Note that the definition of $\hat{\theta}_n$ implies

$$\omega \;\;\in\;\; A^{(n)}_{\hat{\theta}_n(\omega)} \qquad\qquad \forall\,\omega\in\Omega \tag{6.33}$$

Put $\Theta_0 \;:=\; \big\{\theta\in\Theta\;\big|\;\; P'_{\theta_0}\in\mathcal{M}'_\theta\big\}$. Then, the following relations are valid for every $\omega\in\Omega$:

$$P'_{\theta_0}\notin\mathcal{M}'_{\hat{\theta}_n(\omega)} \quad\;\Rightarrow\quad\; \hat{\theta}_n(\omega)\;\in\;\Theta\setminus\Theta_0 \quad\; \overset{(6.33)}{\Rightarrow} \quad\; \omega\;\;\in\;\bigcup_{\theta\,\in\,\Theta\setminus\Theta_0} A_\theta^{(n)}$$

Therefore,

$$U_0^*\Big(\,P'_{\theta_0}\notin\mathcal{M}'_{\hat{\theta}_n}\Big) \quad\le\quad \sum_{\theta\,\in\,\Theta\setminus\Theta_0} U_0^*\big(A_\theta^{(n)}\big) \tag{6.34}$$

For every $\theta\;\in\;\Theta\setminus\Theta_0$, it follows from $P_{\theta_0}\notin\mathcal{M}'_\theta$ that there is a $\varepsilon_\theta>0$ such that

$$\sup_{i\in\{1,\ldots,s\}}\Big(P'_{\theta_0}[f_i]-\overline{P}'_\theta[f_i]\Big)\cdot\|f_i\|^{-1} \quad > \quad \varepsilon_\theta \tag{6.35}$$

Then, for every $\omega\in\Omega$ and for every $\theta\;\in\;\Theta\setminus\Theta_0$,

$$\begin{aligned}
\big\|\mathbb{P}_\omega^{(n)}-\overline{P}'_\theta\big\| \;\;&\ge\;\; \inf_{P'_\theta\in\mathcal{M}'_\theta}\;\sup_{i\in\{1,\ldots,s\}}\Big|\mathbb{P}_\omega^{(n)}[f_i]-P'_\theta[f_i]\Big|\cdot\|f_i\|^{-1}\;\;\ge\\
&\ge\;\sup_{i\in\{1,\ldots,s\}}\;\inf_{P'_\theta\in\mathcal{M}'_\theta}\Big(\mathbb{P}_\omega^{(n)}[f_i]-P'_\theta[f_i]\Big)\cdot\|f_i\|^{-1}\;\;=\\
&=\;\sup_{i\in\{1,\ldots,s\}}\Big(\mathbb{P}_\omega^{(n)}[f_i]-P'_{\theta_0}[f_i]\;+\;P'_{\theta_0}[f_i]-\overline{P}'_\theta[f_i]\Big)\cdot\|f_i\|^{-1}\;\;\ge\\
&\ge\;\inf_{i\in\{1,\ldots,s\}}\Big(\mathbb{P}_\omega^{(n)}[f_i]-P'_{\theta_0}[f_i]\Big)\cdot\|f_i\|^{-1}+\sup_{i\in\{1,\ldots,s\}}\Big(P'_{\theta_0}[f_i]-\overline{P}'_\theta[f_i]\Big)\cdot\|f_i\|^{-1}\\
&\overset{(6.35)}{>}\;\inf_{i\in\{1,\ldots,s\}}\Big(\mathbb{P}_\omega^{(n)}[f_i]-P'_{\theta_0}[f_i]\Big)\cdot\|f_i\|^{-1}\;+\;\varepsilon_\theta
\end{aligned}$$

Put

$$Z_n(\omega) \quad = \quad \left\|\mathbb{P}_\omega^{(n)} - \overline{P}'_{\theta_0}\right\| \; - \inf_{i\in\{1,\dots,s\}} \Big(\mathbb{P}_\omega^{(n)}[f_i] - P'_{\theta_0}[f_i]\Big)\cdot \|f_i\|^{-1} \qquad \forall\,\omega\in\Omega$$

and note that, for every $\theta \in \Theta\setminus\Theta_0$,

$$Z_n(\omega) \;>\; \varepsilon_\theta \qquad \forall\,\omega\in A_\theta^{(n)}$$

Therefore, it follows from (6.34) that

$$U_0^*\Big(\,P'_{\theta_0}\notin \mathcal{M}'_{\hat\theta_n}\Big) \quad\le\quad \sum_{\theta\,\in\,\Theta\setminus\Theta_0} U_0^*\big(Z_n \;>\; \varepsilon_\theta\big) \tag{6.36}$$

Next, Theorem 6.6 and the strong law of large numbers (Hoffmann-Jørgensen, 1994a, § 4.12) yield

$$Z \;\xrightarrow[n\to\infty]{}\; 0 \qquad U_0 \,-\, \text{a.s.}^*$$

According to (van der Vaart and Wellner, 1996, Lemma 1.9.2), U_0 - a.s.* - convergence implies convergence in U_0^* - probability. Hence,

$$U_0^*\Big(\,P'_{\theta_0}\notin \mathcal{M}'_{\hat\theta_n}\Big) \;\overset{(6.36)}{\le}\; \sum_{\theta\,\in\,\Theta\setminus\Theta_0} U_0^*\big(Z_n \;>\; \varepsilon_\theta\big) \;\xrightarrow[n\to\infty]{}\; 0$$

□

6.5 Implementation and application of the estimator

6.5.1 Discretization

As seen in the previous section, it is not necessary to discretize the sample space in order to define the minimum distance estimator based on the total variation norm in a sensible way. Since this is not possible for precise probabilities, going over to imprecise probabilities turns out to be a simplification.

Of course, if we want to calculate the estimator by use of computers, the sample space has to be discretized – at least implicitly. However, it is one of the most striking properties of the above presented minimum distance estimator, that this is only a practical need which is irrelevant for theoretical investigations. In case of precise probabilities, discretization would even be part of the definition of the minimum distance estimator.

Again assume that we have an imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ on a measurable space $(\mathcal{X},\mathcal{A}')$ such that assumptions (6.7) and (6.8) are fulfilled. That is, there is a finite subset $\mathcal{K} = \{f_1,\dots,f_s\}\subset \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ such that, for every $\theta\in\Theta$, the credal set of $\overline{P}'_\theta$ is given by

$$\mathcal{M}'_\theta \quad=\quad \Big\{P'_\theta\in \mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\;\Big|\; P'_\theta[f]\le \overline{P}'_\theta[f] \quad \forall\, f\in\mathcal{K}\Big\}$$

and

$$\overline{P}'_\theta[f]-\underline{P}'_\theta[f] \;>\; 0 \qquad\quad \forall\, f\in\mathcal{K}$$

where $\underline{P}'_\theta$ is the corresponding lower coherent prevision. In addition, assume that Θ is a finite index set.

Then, we are in the situation of Subsection 5.4.2 and we can apply the method of discretizing for some fixed $\varepsilon > 0$ presented therein:

As in (5.40), let $\{A_1, \dots, A_r\}$ be the partition of $\mathcal{X}$ which generates the finite σ-algebra $\mathcal{A}$. For every $i \in \{1, \dots, s\}$, let s_i be the $\mathcal{A}$-simple function which corresponds to f_i according to (5.39). For every $\theta \in \Theta$, let $\overline{Q}_\theta$ be the coherent upper prevision on $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ which corresponds to the credal set

$$\mathcal{N}_\theta \;=\; \big\{\, Q_\theta \in \mathrm{ba}_1^+(\mathcal{X}, \mathcal{A}) \;\big|\; Q_\theta[s_i] \;\leq\; \overline{P}'_\theta[f_i] + \varepsilon_i d_i \quad \forall\, i \in \mathcal{I}_\theta \big\}$$

– confer (5.41). Recall from (5.32) that, due to finiteness of Θ, d_i may be defined to be

$$d_i \;:=\; d_{f_i} \;:=\; \min_{\theta \in \Theta} \overline{P}'_\theta[f_i] - \underline{P}'_\theta[f_i] \;>\; 0$$

for every $i \in \{1, \dots, s\}$.

Now, let

$$x_1\,,\;\dots\,,\;x_n \;\in\; \mathcal{X}$$

be some observations and let

$$\mathbb{P}_x^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\,, \qquad x \;=\; (x_1, \dots, x_n) \;\in\; \mathcal{X}^n\,,$$

be the empirical measure. Then, the minimum distance estimator is

$$\hat{\theta}'_n(x) \;=\; \arg\min_{\theta \in \Theta} \big\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\big\|$$

In order to calculate this estimator by computers, it would be desirable that

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;=\; \inf_{Q_\theta \in \mathcal{N}_\theta} \;\sup_{f \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})} \frac{\big|\,\mathbb{P}_x^{(n)}[f] - Q[f]\,\big|}{\|f\|}$$

was approximately equal to $\big\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\big\|$. Theorem 6.11 below shows that this is true.

Theorem 6.11 *In the setup of the present subsection,*

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;\leq\; \big\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\big\| \;\leq\; \big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| + 2\varepsilon \qquad \forall\, \theta \in \Theta$$

for every $x \in \mathcal{X}^n$.

Proof: Fix any $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ and any $\theta \in \Theta$. Let $\overline{Q}'_\theta$ be the natural extension of $\overline{Q}_\theta$ on $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$. Then,

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\big\| \;=\; \inf_{Q'_\theta \in \mathcal{N}'_\theta} \;\sup_{f' \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')} \frac{\big|\,\mathbb{P}_x^{(n)}[f'] - Q'[f']\,\big|}{\|f'\|}$$

Note that the assumptions in the present subsection guarantee the validity of (5.45). That is,

$$\overline{P}'_\theta[f'] \;\leq\; \overline{Q}'_\theta[f'] \;\leq\; \overline{P}'_\theta[f'] + \varepsilon\big(\sup f' - \inf f'\big) \qquad \forall\, f' \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}')$$

Next, a twofold application of Lemma 6.7 implies

$$\left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta'\right\| \;=\; \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\mathbb{P}_x^{(n)}[f'] - \overline{P}_\theta'[f']}{\|f'\|}$$

and

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| \;=\; \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\mathbb{P}_x^{(n)}[f'] - \overline{Q}_\theta'[f']}{\|f'\|}$$

Hence,

$$\begin{aligned}
\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| \;&\leq\; \left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta'\right\| \;\leq \\
&\leq \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\mathbb{P}_x^{(n)}[f'] - \Big(\overline{Q}_\theta'[f'] - \varepsilon\big(\sup f' - \inf f'\big)\Big)}{\|f'\|} \;\leq \\
&\leq \left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| + \varepsilon \cdot \sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\sup f' - \inf f'}{\|f'\|} \;\leq \\
&\leq \left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| + 2\varepsilon
\end{aligned}$$

Hence,

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| \;\leq\; \left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta'\right\| \;\leq\; \left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| + 2\varepsilon$$

and it only remains to proof

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta'\right\| \;=\; \left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\| \tag{6.37}$$

The inequality " $\geq$ " is trivially fulfilled in (6.37). In order to prove the inequality " $\leq$ " it is enough to show

$$\sup_{f' \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A}')} \frac{\mathbb{P}_x^{(n)}[f'] - \overline{Q}_\theta'[f']}{\|f'\|} \;\leq\; \sup_{f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|} \tag{6.38}$$

according to Lemma 6.7. That is, it only remains to prove (6.38) in the following:

To this end, choose any $a_j \in A_j$ for every element A_j of the partition $\{A_1, \ldots, A_r\}$ of $\mathcal{X}$. Furthermore, put

$$N_j \;=\; \big\{ i \in \{1, \ldots, n\} \;\big|\; x_i \in A_j \big\}$$

and let n_j be the number of elements in N_j for every $j \in \{1, \ldots, r\}$. In addition, put

$$\mathcal{J}_0 \;=\; \big\{ j \in \{1, \ldots, r\} \;\big|\; n_j = 0 \big\} \qquad \text{and} \qquad \mathcal{J}_1 \;=\; \big\{ j \in \{1, \ldots, r\} \;\big|\; n_j > 0 \big\}$$

In particular, this means

$$\{x_1, \ldots, x_n\} \cap A_j \;=\; \emptyset \qquad \forall\, j \in \mathcal{J}_0 \tag{6.39}$$

and

$$\sum_{k=1}^{n} I_{A_j}(x_k) \;=\; n_j \qquad \forall\, j \in \mathcal{J}_1 \tag{6.40}$$

Applying these settings, we can define the map

$$\xi \;:\; \mathcal{L}_\infty(\mathcal{X},\mathcal{A}') \;\to\; \mathcal{L}_\infty(\mathcal{X},\mathcal{A}), \qquad f' \;\mapsto\; \xi(f')$$

where

$$\xi(f') \;=\; \sum_{j\in\mathcal{J}_0} f'(a_j)I_{A_j} + \sum_{j\in\mathcal{J}_1}\Big(\frac{1}{n_j}\sum_{i\in N_j} f'(x_i)\Big)I_{A_j} \tag{6.41}$$

Note that this map is defined well and that $\mathcal{J}_0\cap\mathcal{J}_1=\emptyset$. Then, it is an immediate consequence of the definitions that ξ is linear, positive ($\xi(f')\geq 0\;\forall f'\geq 0$) and normalized ($\xi(I_{\mathcal{X}})=I_{\mathcal{X}}$). Therefore, the adjoint of ξ

$$\rho\;:\;\;\mathrm{ba}(\mathcal{X},\mathcal{A})\;\to\;\mathrm{ba}(\mathcal{X},\mathcal{A}')\,,\qquad \mu\;\mapsto\;\rho(\mu)$$

– defined by $\rho(\mu)[f']=\mu\big[\xi(f')\big]\;\forall\mu\in\mathrm{ba}(\mathcal{X},\mathcal{A})$, $\forall f'\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ – is an ordinary randomization; confer also Proposition 3.11. Especially,

$$\rho(Q)\;\in\;\mathrm{ba}_1^+(\mathcal{X},\mathcal{A}')\qquad\forall\,Q\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A})$$

Note that every $f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ may be written as

$$f\;=\;\sum_{j=1}^r\alpha_jI_{A_j}\;=\;\sum_{j\in\mathcal{J}_0}\alpha_jI_{A_j}+\sum_{j\in\mathcal{J}_1}\alpha_jI_{A_j}$$

for some suitable real numbers $\alpha_1,\dots,\alpha_r$. Then, it follows from the definition (6.41) of ξ that

$$\xi(f)\;=\;f\qquad\forall\,f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A})\subset\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$$

Therefore, $\rho(Q)$ is an extension of Q to a probability charge on $(\mathcal{X},\mathcal{A}')$ for every $Q\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A})$. Since $\overline{Q}_\theta'$ is the natural extension of $\overline{Q}_\theta$, this implies

$$\rho(Q_\theta)\;\in\;\mathcal{N}_\theta'\qquad\forall\,Q_\theta\in\mathcal{N}_\theta \tag{6.42}$$

where $\mathcal{N}_\theta'$ is the credal set of $\overline{Q}_\theta'$. In addition, the following calculation shows that

$$\rho(\mathbb{P}_x^{(n)})\;=\;\mathbb{P}_x^{(n)} \tag{6.43}$$

This is because

$$\begin{aligned}
\mathbb{P}_x^{(n)}\big[\xi(f')\big] \;&=\; \frac{1}{n}\sum_{k=1}^n\Bigg(\sum_{j\in\mathcal{J}_0}f'(a_j)I_{A_j}(x_k)+\sum_{j\in\mathcal{J}_1}\Big(\frac{1}{n_j}\sum_{i\in N_j}f'(x_i)\Big)I_{A_j}(x_k)\Bigg)\\
&\overset{(6.39)}{=}\;\frac{1}{n}\sum_{j\in\mathcal{J}_1}\Bigg(\Big(\frac{1}{n_j}\sum_{i\in N_j}f'(x_i)\Big)\sum_{k=1}^nI_{A_j}(x_k)\Bigg)\;=\\
&\overset{(6.40)}{=}\;\frac{1}{n}\sum_{j\in\mathcal{J}_1}\Bigg(\Big(\frac{1}{n_j}\sum_{i\in N_j}f'(x_i)\Big)\cdot n_j\Bigg)\;=\;\frac{1}{n}\sum_{j\in\mathcal{J}_1}\sum_{i\in N_j}f'(x_i)\;=\\
&=\;\frac{1}{n}\sum_{i=1}^nf'(x_i)\;=\;\mathbb{P}_n^{(n)}[f']\qquad\forall\,f'\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')
\end{aligned}$$

Finally, fix any $f'\in\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')\setminus\{0\}$. Then,

$$\big\|\xi(f')\big\|\;\leq\;\|f'\| \tag{6.44}$$

and

$$\frac{\mathbb{P}_x^{(n)}[f'] - \overline{Q}'_\theta[f']}{\|f'\|} \overset{(6.43)}{=} \inf_{Q'_\theta \in \mathcal{N}'_\theta} \frac{\rho\big(\mathbb{P}_x^{(n)}\big)[f'] - Q'_\theta[f']}{\|f'\|} \leq$$
$$\overset{(6.42)}{\leq} \inf_{Q_\theta \in \mathcal{N}_\theta} \frac{\rho\big(\mathbb{P}_x^{(n)}\big)[f'] - \rho(Q_\theta)[f']}{\|f'\|} = \inf_{Q_\theta \in \mathcal{N}_\theta} \frac{\mathbb{P}_x^{(n)}\big[\xi(f')\big] - Q_\theta\big[\xi(f')\big]}{\|f'\|}$$

If $\xi(f') = 0$, this implies

$$\frac{\mathbb{P}_x^{(n)}[f'] - \overline{Q}'_\theta[f']}{\|f'\|} \leq 0 = \frac{\mathbb{P}_x^{(n)}[I_\mathcal{X}] - \overline{Q}_\theta[I_\mathcal{X}]}{\|I_\mathcal{X}\|} \leq \sup_{f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|}$$

and, if $\xi(f') \neq 0$, this implies

$$\frac{\mathbb{P}_x^{(n)}[f'] - \overline{Q}'_\theta[f']}{\|f'\|} \leq \inf_{Q_\theta \in \mathcal{N}_\theta} \frac{\mathbb{P}_x^{(n)}\big[\xi(f')\big] - Q_\theta\big[\xi(f')\big]}{\|f'\|} \leq$$
$$\overset{(6.44)}{\leq} \inf_{Q_\theta \in \mathcal{N}_\theta} \frac{\mathbb{P}_x^{(n)}\big[\xi(f')\big] - Q_\theta\big[\xi(f')\big]}{\big\|\xi(f')\big\|} = \frac{\mathbb{P}_x^{(n)}\big[\xi(f')\big] - \overline{Q}_\theta\big[\xi(f')\big]}{\big\|\xi(f')\big\|} \leq$$
$$\leq \sup_{f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{A})} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|}$$

Therefore, (6.38) follows. □

So, Theorem 6.11 justifies the following proceeding:

- An application of Subsection 5.4.2 turns the imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A}')$ into a discretized model $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A})$.
- Calculating $\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\|$ yields an approximation of $\big\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\big\|$.

Although $(\mathcal{X},\mathcal{A})$ is a finite space, it is still an issue how to calculate

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| = \inf_{Q_\theta \in \mathcal{N}_\theta} \sup_{f \in \mathcal{L}(\mathcal{X},\mathcal{A})} \frac{\big|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\big|}{\|f\|} \tag{6.45}$$

since $\mathcal{N}_\theta$ and $\mathcal{L}(\mathcal{X},\mathcal{A})$ still are infinite spaces. As done in Troffaes (2008), it would – in principle – be possible to discretize these sets as well. However, this would not lead to an applicable method because of exceedingly high computational costs. Instead, the value of the distance can be calculated by means of linear programming as shown in the following subsection.

Since we are not really interested in $\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\|$ but in

$$\arg\inf_{\theta\in\Theta} \big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\|$$

the following question arises: Can this $\arg\inf$ be calculated by calculating

$$\arg\inf_{\theta\in\Theta} \inf_{Q_\theta \in \mathcal{N}_\theta} \sup_{f \in \mathcal{K}} \frac{\big|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\big|}{\|f\|} \tag{6.46}$$

where the infinite set $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ is replaced by the finite set $\mathcal{K}$? If this was possible, calculating the minimum distance estimator would become much more easier. However, this is not possible as can be seen from the following example. So, we cannot avoid calculating (6.45) in this simple way.

Example 6.12 *Take* $\mathcal{X}=\{1,2,3,4\}$ *and assume that we have* $n=4$ *observations:*

$$x_1=1\,,\quad x_2=2\,,\quad x_3=3\,,\quad x_4=4$$

That is $\mathbb{P}_x^{(4)}=\frac{1}{4}\delta_1+\frac{1}{4}\delta_2+\frac{1}{4}\delta_3+\frac{1}{4}\delta_4\,.$ *Furtheremore, take* $\mathcal{K}=\{I_{\{1\}},\ I_{\{2\}},\ I_{\{3\}}\}$ *and consider the coherent upper previsions* $\overline{Q}_0$ *and* $\overline{Q}_1$ *on* $\mathcal{K}$ *defined by*

$$\overline{Q}_0[I_{\{1\}}]\ =\ \tfrac{1}{4}-3\alpha\,,\quad \overline{Q}_0[I_{\{2\}}]\ =\ \tfrac{1}{4}-3\alpha\,,\quad \overline{Q}_0[I_{\{3\}}]\ =\ \tfrac{1}{4}+\alpha$$

and

$$\overline{Q}_1[I_{\{1\}}]\ =\ \tfrac{1}{4}-4\alpha\,,\quad \overline{Q}_2[I_{\{2\}}]\ =\ \tfrac{1}{4}+4\alpha\,,\quad \overline{Q}_3[I_{\{3\}}]\ =\ \tfrac{1}{4}-\alpha$$

for any real number $0<\alpha<\frac{1}{12}\,.$ *That is* $\Theta=\{0,1\}\,.$ *Then, the* arg inf *in (6.46) would be* $\theta=0\,.$ *But,*

$$\big\|\mathbb{P}_x^{(4)}-\overline{Q}_0\big\|\ =\ 12\alpha\ >\ 10\alpha\ =\ \big\|\mathbb{P}_x^{(4)}-\overline{Q}_1\big\|$$

where $\big\|\mathbb{P}_x^{(4)}-\overline{Q}_0\big\|$ *and* $\big\|\mathbb{P}_x^{(4)}-\overline{Q}_0\big\|$ *can be calculated according to the methods presented in the following subsection.*

6.5.2 Calculation by linear programming

This subsection is concerned with the question how to calculate

$$\big\|\mathbb{P}_x^{(n)}-\overline{Q}_\theta\big\|\ =\ \inf_{Q_\theta\in\mathcal{N}_\theta}\ \sup_{f\in\mathcal{L}(\mathcal{X},\mathcal{A})}\frac{\big|\mathbb{P}_x^{(n)}[f]-Q_\theta[f]\big|}{\|f\|}\tag{6.47}$$

where $\overline{Q}_\theta$ is the coherent upper prevision with credal set

$$\mathcal{N}_\theta\ =\ \big\{\,Q_\theta\in\mathrm{ba}_1^+(\mathcal{X},\mathcal{A})\ \big|\ Q_\theta[s_k]\ \le\ \overline{P}'_\theta[f_k]+\varepsilon_k d_k\quad\forall\,k\in\mathcal{I}_\theta\big\}$$

– confer (5.41) – on the finite sample space $(\mathcal{X},\mathcal{A})$. Furthermore, $x=(x_1,\dots,x_n)\in\mathcal{X}^n$ is the vector of all observations

$$x_1\,,\ \dots\ ,\ x_n$$

That is, we have to minimize the convex function

$$\mathcal{N}_\theta\ \to\ \mathbb{R}\,,\qquad Q_\theta\ \mapsto\ \sup_{f\in\mathcal{L}(\mathcal{X},\mathcal{A})}\frac{\big|\mathbb{P}_x^{(n)}[f]-Q_\theta[f]\big|}{\|f\|}$$

However, we do not need any results of convex optimization in order to do this. Linear optimization is enough – as a matter of fact, the value in (6.47) may be calculates by one single linear program.

To this end, recall the following definitions from the proof of Theorem 6.11:

Let again $\{A_1,\dots,A_r\}\subset\mathcal{A}$ be the partition of $\mathcal{X}$ such that every $A\in\mathcal{A}$ is the union of some elements of $\{A_1,\dots,A_r\}$. For every $j\in\{1,\dots,r\}$, choose any $a_j\in A_j$, put

$$N_j\ =\ \big\{i\in\{1,\dots,n\}\ \big|\ x_i\in A_j\big\}$$

and let n_j be the number of elements in N_j . Furthermore,

$$\mathcal{J}_0 \;=\; \big\{ j \in \{1,\dots,r\} \;\big|\; n_j = 0 \big\} \qquad \text{and} \qquad \mathcal{J}_1 \;=\; \big\{ j \in \{1,\dots,r\} \;\big|\; n_j > 0 \big\}$$

In addition, put

$$\hat{\varepsilon}_k \;:=\; \varepsilon_k d_k \qquad \forall\, k \in \{1,\dots,s\} \tag{6.48}$$

Now, consider the following linear program:

$$\sum_{j\in\mathcal{J}_1} q_j - \gamma_j \quad\longrightarrow\quad \text{max!} \tag{6.49}$$

where

$$\sum_{j=1}^{r} q_j \;=\; 1 \tag{6.50}$$

and

$$\sum_{j=1}^{r} q_j s_k(a_j) \;\leq\; \overline{P}'_{\theta}[f_k] + \hat{\varepsilon}_k \qquad \forall\, k \in \mathcal{I}_\theta \tag{6.51}$$

and

$$q_j - \gamma_j \;\leq\; \frac{n_j}{n} \qquad \forall\, j \in \mathcal{J}_1 \tag{6.52}$$

for the variables

$$(q_1,\dots,q_r) \;\in\; \mathbb{R}^r\,, \qquad q_j \geq 0 \quad \forall\, j \in \{1,\dots,r\}\,, \tag{6.53}$$

and

$$(\gamma_j)_{j\in\mathcal{J}_1} \;\subset\; \mathbb{R}\,, \qquad \gamma_j \geq 0 \quad \forall\, j \in \mathcal{J}_1 \tag{6.54}$$

Let β_θ be the optimal value of the above linear program. Then, Proposition 6.13 shows that

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;=\; 2\cdot\big(1-\beta_\theta\big)$$

Hence, it is, in fact, enough to solve one single linear program in order to obtain the distance $\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\|$. Of course, this was useless in applications if this linear program would tend to be unsolvable because of exceedingly high computational costs. So let us take a closer look on the size of the above linear program:
Since the number of elements in $\mathcal{J}_1$ is not larger than $r \wedge n \;=\; \min\{r,n\}$, we have the following upper bounds:

Number of variables: $r \,+\, r\wedge n$

Number of inequalities: $2 \,+\, \sharp(\mathcal{K}_\theta) \,+\, r\wedge n$

According to Subsection 5.4.2, r can – in general – exceed beyond all reasonable bounds (e.g. $r = 10^{300}$) but will stay within a reasonable order of magnitude in most applications; confer e.g. Proposition 5.16.
Though the number n of observations may be very large, it will hardly reach astronomical orders of magnitude in real applications.
The size of the number of elements in $\mathcal{K}_\theta$ (i.e. the number of elements in $\mathcal{I}_\theta$) will usually be negligible.

In the example presented in the end of Subsection 5.4.2, we have the following upper bounds if the assumptions of Proposition 5.16 are fulfilled:

	$n = 100$	$n = 1000$	$n = 10000$	$n \geq 50000$
Number of variables:	40140	41040	50040	80180
Number of inequalities:	112	1012	10012	40052

Solving such linear programs will usually be possible within a reasonable time frame – the more so as a large number of observation leads to a sparse matrix of coefficients in the linear program since nearly all inequalities are given by (6.52) then. In case of such sparse matrices, algorithms are available which can solve huge problems very efficiently.
Furthermore, note that the last column of the above table gives bounds which do not depend on the number of observations any more. That is, the size of the linear program stays constant if the number of observations grows to infinity. If r is not too large, then the estimator can be calculated for any number of observations.

A very large r will usually result from small values $\varepsilon_k d_k$. However, as already stated on page 149, $\overline{P}'_\theta$ cannot be specified so accurately in applications that too small values $\varepsilon_k d_k$ are meaningful. Such small values $\varepsilon_k d_k$ indicates that the imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ is in danger of being instable – confer Section 5.2. This justifies the alternate proceeding presented on page 149 where (5.52) is relaxed to (5.53). Firstly, this means that we have to take

$$\hat{\varepsilon}_k \;:=\; \varepsilon\big(\sup f_k - \inf f_k\big)$$

instead of $\hat{\varepsilon}_k = \varepsilon_k d_k$ in (6.51) of the linear program. Secondly, this means that M is not chosen in order to fulfill (5.37) in the discretization method presented in Subsection 5.4.2. Instead, M has to be chosen so that

$$M - 1 \;<\; \frac{1}{\varepsilon} \;\leq\; M$$

Then, analog to Proposition 5.51, an upper bound on the size r would be

$$4 \cdot s \cdot \left(1 + \frac{1}{\varepsilon}\right)$$

Hence, we end up with a linear program of a very small size which will nearly always be solvable. But, by doing this, it is not guaranteed that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ still is an approximation of $\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\|$. On the other hand, this proceeding is more conservative and, if small changes[13] of $\overline{P}'_\theta$ have large effects on $\|\mathbb{P}_x^{(n)} - \overline{P}'_\theta\|$, it is a good idea to be more conservative because this may save from arbitrary results.

[13] That is, changes which are small but not as small as $\varepsilon_k d_k$ which is assumed to be very small here.

The following proposition says that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ can indeed be calculated by solving the linear program given by (6.49) – (6.54):

Proposition 6.13 *Let β_θ be the optimal value of the linear program given by (6.49) – (6.54). Then,*

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;=\; 2\cdot\big(1-\beta_\theta\big)$$

Proof:

[1] Firstly, it is shown that

$$\|\mathbb{P}_x^{(n)} - Q\| \;=\; 2\cdot\sum_{j\in\mathcal{J}_1}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ \qquad \forall\, Q\in\mathcal{N}_\theta \tag{6.55}$$

To this end, fix any $Q\in\mathcal{N}_\theta$ and note that – due to finiteness of $\mathcal{A}$ – the total variation distance is equal to

$$\|\mathbb{P}_x^{(n)} - Q\| \;=\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\big| \tag{6.56}$$

Since $\{A_1,\dots,A_r\}$ is a partition of $\mathcal{X}$, we have

$$\begin{aligned} 0 \;&=\; \mathbb{P}_x^{(n)}(\mathcal{X}) - Q(\mathcal{X}) \;=\; \sum_{j=1}^{r}\mathbb{P}_x^{(n)}(A_j) - Q(A_j) \;= \\ &=\; \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ - \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^- \end{aligned}$$

Hence,

$$\begin{aligned} \|\mathbb{P}_x^{(n)} - Q\| \;&\overset{(6.56)}{=}\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\big| \;= \\ &=\; \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ + \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^- \;= \\ &=\; 2\cdot\sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ \end{aligned}$$

Then, (6.55) follows from the following assertions:

$$j\notin\mathcal{J}_1 \quad\Rightarrow\quad \mathbb{P}_x^{(n)}(A_j)=0 \quad\Rightarrow\quad \Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ \;=\; 0$$

[2] Secondly, it is shown that, for every $Q\in\mathcal{N}_\theta$ and every $j\in\mathcal{J}_1$,

$$\Big(\mathbb{P}_x^{(n)}(A_j) - Q(A_j)\Big)^+ \;=\; \inf_{\gamma_j\in\Gamma_j(Q)}\mathbb{P}_x^{(n)}(A_j) - Q(A_j) + \gamma_j \tag{6.57}$$

where

$$\Gamma_j(Q) \;:=\; \big\{\gamma_j\in\mathbb{R}\;\big|\;\gamma_j\geq 0\,,\quad Q(A_j)-\gamma_j\;\leq\;\mathbb{P}_x^{(n)}(A_j)\big\}$$

In case of $\mathbb{P}_x^{(n)}(A_j) > Q(A_j)$, it is easy to see that the infimum is attained in $\tilde{\gamma}_i = 0 \in \Gamma_j(Q)$ and, therefore, (6.57) is fulfilled.

In case of $\mathbb{P}_x^{(n)}(A_j) > Q(A_j)$, it is easy to see that the infimum is attained in $\tilde{\gamma}_i = Q(A_j) - \mathbb{P}_x^{(n)}(A_j) \;\in\; \Gamma_j(Q)$ and, therefore, (6.57) is again fulfilled.

[3] Finally, put

$$\mathbb{M} \;=\; \Big\{(Q,\gamma) \in \mathcal{N}_\theta \times \mathbb{R}^{\sharp(\mathcal{J}_1)} \;\Big|\; \gamma = (\gamma_j)_{j\in\mathcal{J}_1}\,, \quad \gamma_j \in \Gamma_j(Q) \quad \forall\, j \in \mathcal{J}_1\Big\}$$

Then, it follows from part [1] and part [2] that

$$\begin{aligned}
\inf_{Q\in\mathcal{N}_\theta} \|\mathbb{P}_x^{(n)} - Q\| \;&\overset{(6.55),(6.57)}{=}\; \inf_{Q\in\mathcal{N}_\theta} 2\cdot\sum_{j\in\mathcal{J}_1} \inf_{\gamma_j\in\Gamma_j(Q)} \mathbb{P}_x^{(n)}(A_j) - Q(A_j) + \gamma_j \\
&= \; 2\cdot \inf_{Q\in\mathcal{N}_\theta} \inf_{\substack{\gamma_j\in\Gamma_j(Q)\\ \forall\, j\in\mathcal{J}_1}} \sum_{j\in\mathcal{J}_1} \mathbb{P}_x^{(n)}(A_j) - Q(A_j) + \gamma_j \; = \\
&= \; 2\cdot \inf_{(Q,\gamma)\in\mathbb{M}} \sum_{j\in\mathcal{J}_1} \mathbb{P}_x^{(n)}(A_j) - Q(A_j) + \gamma_j \qquad (6.58)
\end{aligned}$$

The definition of $\mathcal{J}_1$ implies that

$$\sum_{j\in\mathcal{J}_1} \mathbb{P}_x^{(n)}(A_j) \;=\; 1$$

Hence,

$$\begin{aligned}
\inf_{Q\in\mathcal{N}_\theta} \|\mathbb{P}_x^{(n)} - Q\| \;&\overset{(6.58)}{=}\; 2\cdot\Big(1 + \inf_{(Q,\gamma)\in\mathbb{M}} \sum_{j\in\mathcal{J}_1} -Q(A_j) + \gamma_j\Big) \;= \\
&= \; 2\cdot\Big(1 - \sup_{(Q,\gamma)\in\mathbb{M}} \sum_{j\in\mathcal{J}_1} Q(A_j) - \gamma_j\Big)
\end{aligned}$$

For every $j \in \{1,\dots,r\}$, identify $Q(A_j)$ with the variable q_j in the linear program. Then, it follows from the definitions of $\mathcal{N}_\theta$ and $\mathbb{M}$ that

$$\sup_{(Q,\gamma)\in\mathbb{M}} \sum_{j\in\mathcal{J}_1} Q(A_j) - \gamma_j \;=\; \beta_\theta$$

and, therefore,

$$\inf_{Q\in\mathcal{N}_\theta} \|\mathbb{P}_x^{(n)} - Q\| \;=\; 2\cdot\big(1-\beta_\theta\big)$$

□

6.6 Simulation study

In order to demonstrate the applicability of the minimum distance estimator, this section presents a simulation study consisting of three different models.

As stated in the introductory Section 6.1, the estimator is based on a rather simple idea. This conceptual simplicity enables many possible applications as can be seen in the following. Model 1 of the simulation study is a first example which shows that the minimum distance estimator can also be used for large sample sizes and that going over to an imprecise model does not necessarily mean to loose much efficiency. Model 2 is an application based on normal distributions which is motivated by the popular chi-square test. Model 3 shows how the estimator can be used for linear regression where the (imprecisely known) error distribution does not need to be symmetric.

6.6.1 Model 1: A first example

Model 1 is intended to demonstrate two aspects of the proposed estimator:
Firstly, the estimator can really be calculated even for large numbers of observations – something which is not self-evident for imprecise probabilities. In the simulation study, the estimator is applied for the following sample sizes:

$$n = 30\,, \qquad n = 100\,, \qquad n = 1000\,, \qquad n = 10000$$

For each number of observations, the estimator is evaluated 500 times.
Secondly, the estimator can provide good results even though it is developed for the rather large imprecise models given by (6.7). In order to demonstrate this, the imprecise Model 1 contains a nice precise parametric model so that the estimator can be compared with a maximum likelihood estimator. While the maximum likelihood estimator is applied by using complete knowledge of the precise parametric model, our minimum distance estimator is only based on the knowledge of a large imprecise model. Since the simulated data exactly stem from the ideal parametric model, this is a rather unequal situation which favors the maximum likelihood estimator and, therefore, the maximum likelihood estimator should clearly beat our estimator. Nevertheless, the performance of our estimator appears to be almost as good as the one of the maximum likelihood estimator in the simulation study. In this way, it can be seen that going over to a large imprecise model does not necessarily mean to loose a lot of efficiency even if the ideal parametric model was precisely true. Model 2 and Model 3 below demonstrates what happens in more realistic situations where data do not precisely stem from such an ideal parametric model.

Here is a detailed description of Model 1:

The sample space is $(\mathcal{X}, \mathcal{A}')$ where $\mathcal{X}$ is equal to $[0,1]$ and $\mathcal{A}'$ is the Borel-σ-algebra. The precise parametric model $(P'_\theta)_{\theta\in\Theta}$ is given by

$$dP'_\theta \;=\; p'_\theta \, d\lambda\,, \qquad \theta \,\in\, \Theta := [-2, 2]$$

where the Lebesgue-densities p'_θ are

$$p'_\theta(x) \;=\; 1 + \theta\big(x - 0.5\big) I_{[0,0.5]}(x) + \theta\big(0.75 - x\big) I_{(0.5,1]}(x) \qquad \forall\, x \in [0,1]$$

Despite of this confusing formula, the densities p'_θ are very simple and natural as can be seen from Figure 6.1 which shows the graphs of p'_θ for $\theta = 0$ (this is the uniform distribution), $\theta = 1.5$ and $\theta = -0.5$. In order to define the imprecise model, the parameter set Θ is discretized as suggested in Subsection 6.2.2 by putting

$$\Theta_0 \;:=\; \big\{ \theta \in \Theta \,\big|\; \theta \;=\; -2 + 0.1 \cdot k - 0.05 \;,\;\; k \in \{1, \dots, 40\} \big\}$$

That is, $\theta_0 \in \Theta$ corresponds to the interval $(\theta_0 - 0.05\,, \theta_0 + 0.05]$ with center θ_0 . In accordance with Assumption (6.7), the imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta_0}$ is given by credal sets

$$\mathcal{M}'_\theta \;=\; \big\{ Q'_\theta \;\big|\; Q'_\theta[f] \,\leq\, \overline{P}'_\theta[f] \;\, \forall\, f \in \mathcal{K} \big\} \qquad \forall\, \theta \in \Theta_0$$

Here, $\mathcal{K}$ is the finite set

$$\mathcal{K} \;=\; \big\{ f_1, \dots, f_{10} \big\}$$

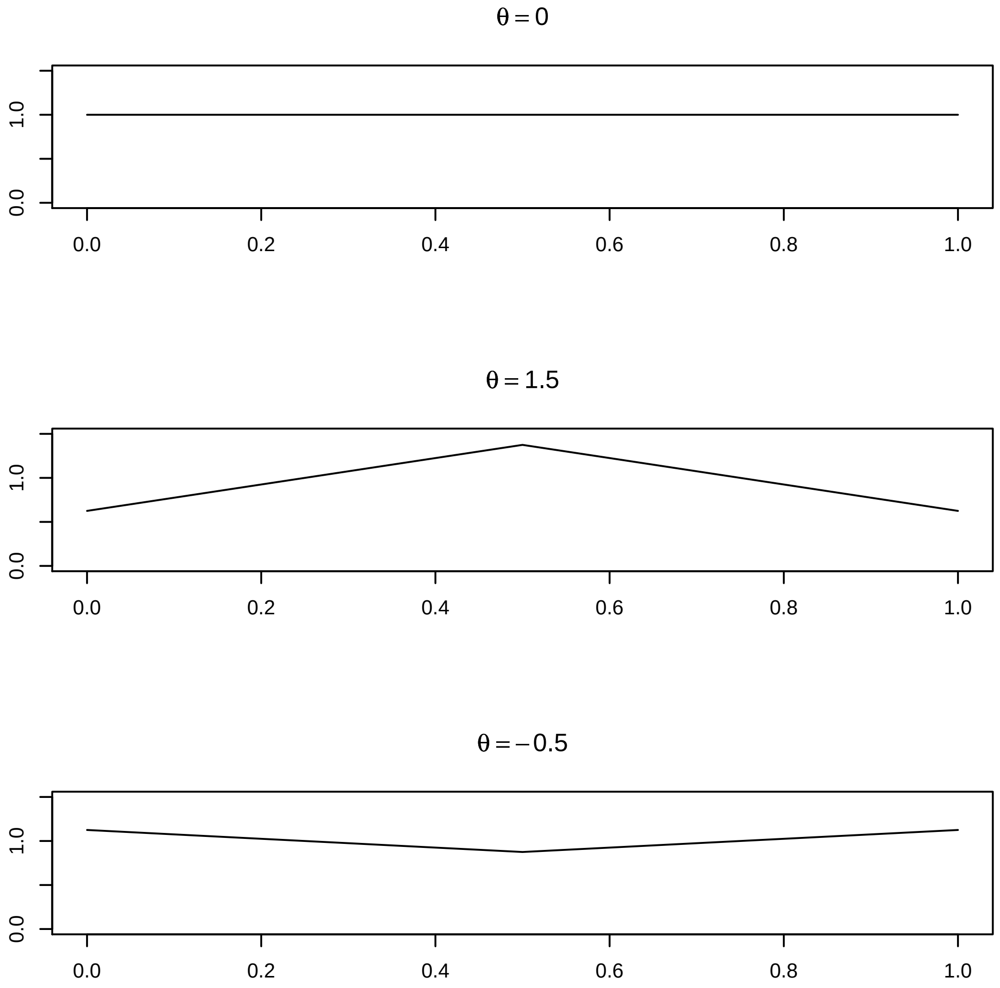

Figure 6.1: Graphs of p'_θ for $\theta = 0$ (the uniform distribution), $\theta = 1.5$ and $\theta = -0.5$ in Model 1

which consists of the (rather arbitrarily chosen) functions $f_j \,:\, [0,1] \to \mathbb{R}\,,\ x \mapsto f_j(x)$ given by

$$f_1(x) \,=\, x\,, \qquad f_2(x) \,=\, 1-x\,, \qquad f_3(x) \,=\, x^2\,, \qquad f_4(x) \,=\, x^3\,,$$

$$f_5(x) \,=\, I_{[\frac{1}{4},\frac{3}{4}]}(x)\,, \qquad f_6(x) \,=\, I_{[0,\frac{1}{4}]}(x)\,, \qquad f_7(x) \,=\, I_{[\frac{3}{4},1]}(x)\,,$$

$$f_8(x) \,=\, \sqrt{x}\,, \qquad f_9(x) \,=\, x + \tfrac{1}{2}I_{[\frac{1}{4},\frac{1}{2}]}(x)\,, \qquad f_{10}(x) \,=\, 4(x-x^2)$$

and the upper previsions on these functions are defined by

$$\overline{P}'_{\theta_0}[f_j] \;=\; \sup_{\theta\in[\theta_0-0.05\,,\,\theta_0+0.05]} \int_0^1 f_j(x)p'_\theta(x)\,\lambda(dx) \qquad \forall\, j \in \{1,\dots,10\} \tag{6.59}$$

for every $\theta_0 \in \Theta_0$. Though there are some similarities, note that this imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta_0}$ is *not* parametrically generated in the sense of Definition 3.24 because (6.59) is not valid for all functions in $\mathcal{L}_\infty(\mathcal{X},\mathcal{A}')$ but only for the small number of functions in $\mathcal{K}$. This makes a great difference; as a consequence, $\mathcal{M}'_\theta$ is a very large credal set while parametrically generated previsions have rather small credal sets.

As already mentioned above, the simulation study consists of 500 runs with different sample sizes $n = 30\,,\ 100\,,\ 1000$ and 10000. The data $x_1,\dots,x_n$ are independent identically distributed by the ideal probability measure P'_0 which is equal to the uniform distribution $\mathrm{Unif}\big([0,1]\big)$. That is, we have

$$X_1,\dots,X_n \ \sim_{\text{i.i.d.}}\ P'_0 \qquad \text{where} \qquad P'_0 \,=\, \mathrm{Unif}\big([0,1]\big)$$

and $\theta = 0$ is the true parameter which has to be estimated.

For the estimation, the proposed minimum distance estimator and the maximum likelihood estimator are applied.

The actual calculation of the proposed minimum distance estimator slightly differs from the one presented in Subsection 6.5.2: As suggested on page 149 and page 182, $\hat{\varepsilon}_k = \varepsilon_k d_k$ is replaced by a larger value in the discretization. This is justified by the fact that too small values $\hat{\varepsilon}_k = \varepsilon_k d_k$ are not meaningful in applications since the values $\overline{P}'_\theta$ cannot be specified with such an accuracy. Instead of $\hat{\varepsilon}_k = \varepsilon_k d_k$, the value $\varepsilon = 0.0005$ has been taken which is still quite small. Recall that going over to this value corresponds to a more cautious proceeding.

The maximum likelihood estimator is defined to be

$$\hat{\theta}_{n,\text{MaxLikelihood}}(x_1,\dots,x_n) \;=\; \arg\max_{\theta\in[-2,2]} \prod_{i=1}^n p_\theta(x_i)$$

Note that – due to the discretization of Θ – our minimum distance estimator does not specify a precise value θ as an estimation but an interval $[\theta_0-0.05, \theta_0+0.05]$. In order to compare the results between both estimators, these intervals $[\theta_0-0.05, \theta_0+0.05]$ are recorded by their center θ_0.

Table 6.1 shows the empirical mean squared error (MSE)

$$\frac{1}{500}\sum_{j=1}^{500} \big(\hat{\theta}_n^{(j)} \,-\, 0\big)^2$$

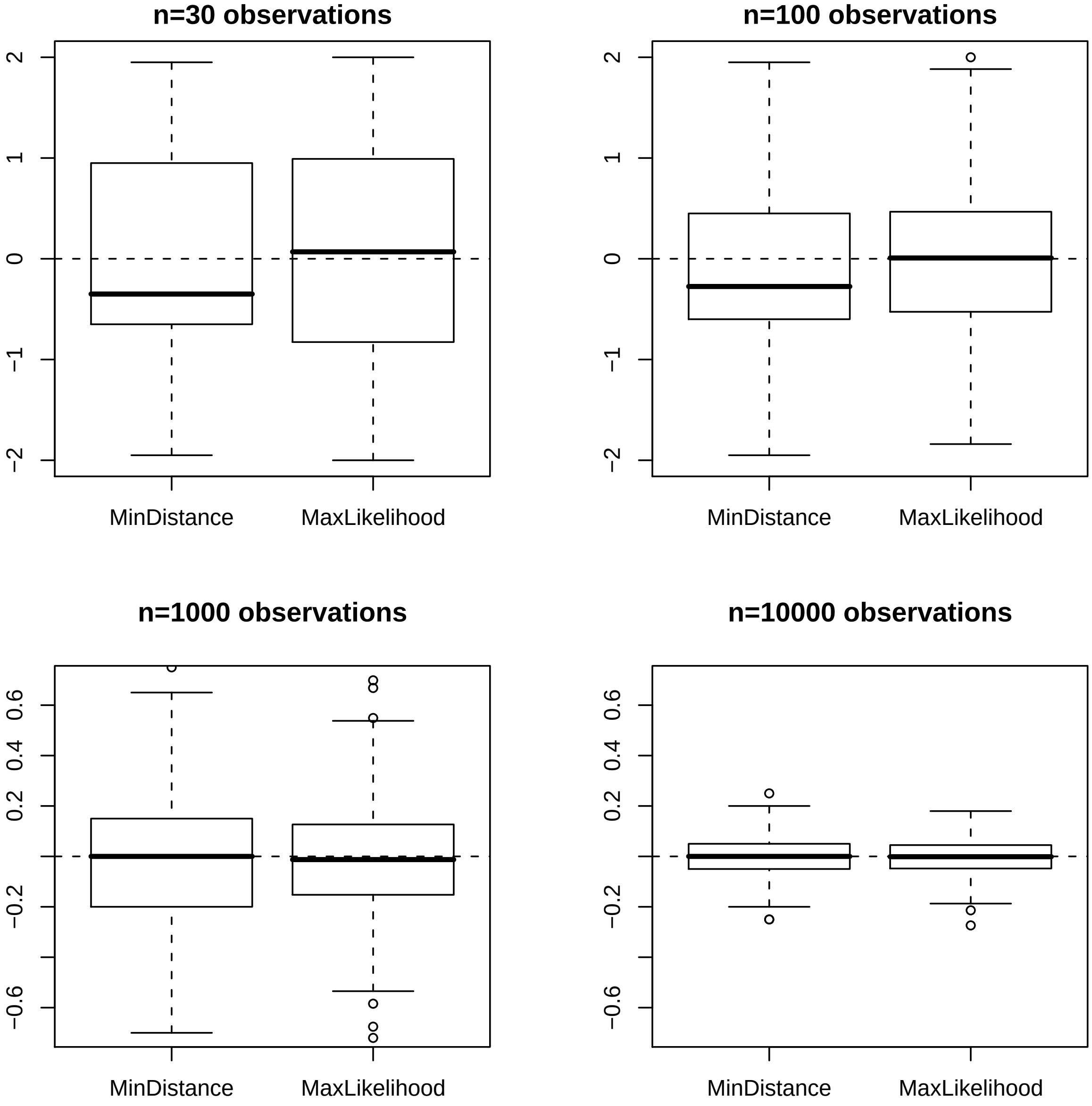

Figure 6.2: Boxplots of the estimations obtained in 500 runs for each number of observations in Model 1

n	MinDistance	MaxLikelihood
30	1.29943	1.35598
100	0.59675	0.49674
1000	0.06753	0.04692
10000	0.00711	0.00482

Table 6.1: Empirical mean squared error calculated over the estimations obtained in 500 runs for each number of observations in Model 1

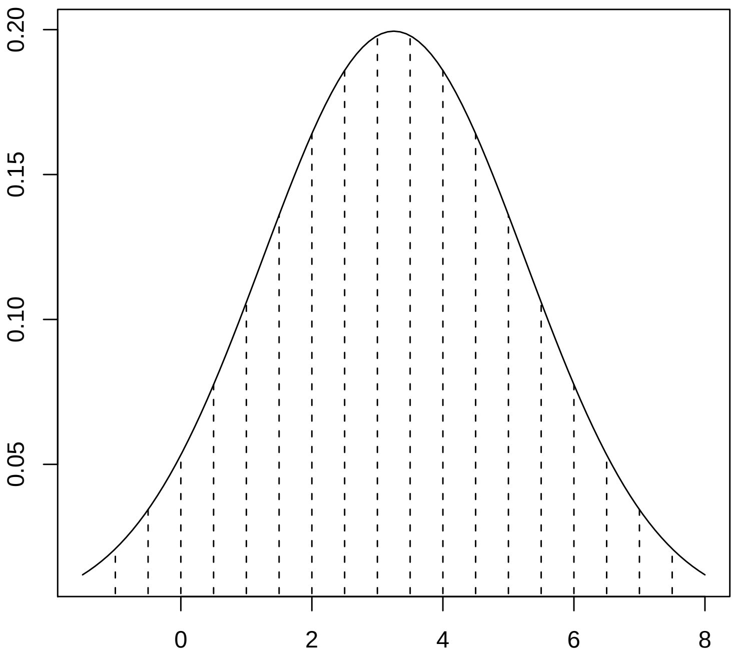

Figure 6.3: Densitiy of the normal distribution $\mathcal{N}(3.25, 2)$; the sample space is divided into segments of width 0.5 .

of the estimations $\hat{\theta}_n^{(j)}$ calculated over all runs $j = 1, \ldots, 500$ for the proposed minimum distance estimator (MinDistance) and the classical maximum likelihood estimator (MaxLikelihood); these values are similar for both estimators. Figure 6.2 shows the boxplots of the estimations. These results demonstrate that, in Model 1, the maximum likelihood estimator is not much better than the minimum distance estimator even though the unequal situation of Model 1 highly privilege the maximum likelihood estimator as explained above.

6.6.2 Model 2: Approximate normal distributions

After considering the rather artificial Model 1 which should demonstrate that the estimator really works, we may turn over to a practical example now. Many statistical evaluations are based on the assumption that the data stem from a normal distribution. Though it is not possible to statistically assure the validity of this assumption, it is often tried to do this by a chi-square test. In order to do this, the sample space $\mathcal{X} = \mathbb{R}$ is divided into segments as shown in Figure 6.3. Since the chi-square test only takes the probabilities of such segments into account, the test is far away from covering all aspects of the normal distribution. Therefore, this situation does not cope with the strict assumption of a precise normal distribution but exactly corresponds to partially determinated F-probabilities (cf. Subsection 2.4.5). This motivates the following definition of Model 2.

The sample space $(\mathcal{X}, \mathcal{A}')$ is equal to $(\mathbb{R}, \mathfrak{B})$. It is divided by

$$a_0 = -10\,,\ \ a_1 = -9.5\,,\ \ a_2 = -9\,,\ \ \ldots\,,\ \ a_k = -10 + k \cdot 0.5\,,\ \ \ldots\,,\ \ a_{40} = 10$$

into the segments

$$(-\infty, a_0)\,, \quad (a_0, a_1]\,, \quad (a_1, a_2]\,, \quad \ldots\,, \quad (a_{39}, a_{40}]\,, \quad (a_{40}, \infty)$$

That is, we consider the set of functions

$$\mathcal{K} \;=\; \big\{f_0, f_1, f_2, \ldots, f_{40}, f_{41}\big\}$$

where

$$f_0 \;=\; I_{(-\infty, a_0]}\,, \qquad f_{41} \;=\; I_{(a_{40}, \infty)}$$

and

$$f_k \;=\; I_{(a_{k-1}, a_k]} \qquad \forall\, k \in \{1, \ldots, 40\}$$

According to the above motivation, we want to deal with a family of normal distributions on $(\mathbb{R}, \mathfrak{B})$

$$P'_\theta \;=\; \mathcal{N}(\mu, \sigma^2) \qquad \text{where} \quad \theta \;=\; (\mu, \sigma) \quad \text{for} \quad -5 \,\le\, \mu \,\le\, 5\,, \quad 0.4 \,\le\, \sigma \,\le\, 2$$

That is, the index set is[14]

$$\Theta \;=\; \Theta^{(1)} \times \Theta^{(2)} \;=\; [-5, 5] \times [0.4, 2]$$

For the definition of the imprecise model, Θ is again discretized:

$$\Theta_0 \;=\; \left\{ (\mu_0, \sigma_0) \in \Theta \;\middle|\; \begin{array}{l} \mu_0 \;=\; -5 + 0.2 \cdot k_1 - 0.1\,,\ k_1 \in \{1, \ldots, 50\}\,, \\ \sigma_0 \;=\; 0.4 + 0.2 \cdot k_2 - 0.1\,,\ k_2 \in \{1, \ldots, 8\} \end{array} \right\}$$

That is, (μ_0, σ_0) corresponds to the rectangle $(\mu_0 - 0.1, \mu_0 + 0.1] \times (\sigma_0 - 0.1, \sigma_0 + 0.1]$ with center (μ_0, σ_0). Based on this discretization and the normal distributions, we can define the following upper previsions for the segments of the sample space:

$$\overline{P}'_{\theta_0}[f_j] \;=\; (1 - 0.02) \cdot \sup_{\substack{\mu \in (\mu_0 - 0.1, \mu_0 + 0.1] \\ \sigma \in (\sigma_0 - 0.1, \sigma_0 + 0.1]}} \int_{\mathbb{R}} f_j \; d\mathcal{N}(\mu, \sigma^2) \;+\; 0.02$$

for every $j \in \{1, \ldots, 40\}$ and $(\mu_0, \sigma_0) \in \Theta_0$. The value 0.02 leads to more imprecision in the imprecise model – that is, to a more cautious proceeding. Roughly speaking, 0.02 can be interpreted as the probability that the data stem from any distribution which can be totally different from normal distributions. This proceeding is very similar to the use of the contamination neighborhoods in robust statistics. The credal sets of the coherent upper previsions are given by

$$\mathcal{M}'_{\theta_0} \;=\; \big\{Q'_\theta \;\big|\; Q'_\theta[f_j] \,\le\, \overline{P}'_{\theta_0}[f_j] \ \ \forall\, j \in \{1, \ldots, 40\}\big\} \qquad \forall\, \theta_0 \in \Theta_0 \tag{6.60}$$

[14] This considerable restriction of the index set is not crucial for the calculation of the minimum distance estimator because its implementation guaranties that the computational effort increases at most linearly with the size of the index set. However, this makes the simulation study easier which consists of two times 500 runs.

Note that these credal sets are much larger than contamination neighborhoods of normal distributions with radius 0.02 because, in (6.60), only the probabilities of the above segments are associated with normal distributions.

The simulation study has two parts where each part consists of 500 runs with sample size $n = 100$. In the first part, the data $x_1, \ldots, x_n$ are independent, identically distributed by the ideal probability measure $P'_\theta = \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma) = (-4, 1.4)$. That is,

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathcal{N}(\mu, \sigma^2)$$

where $\mu = -4$ and $\sigma = 1.4$ are the true parameters which have to be estimated. In the second part, the data $x_1, \ldots, x_n$ are independent, identically distributed by the probability measure $P'_\theta = 0.7 \cdot \mathcal{N}(\mu, \sigma^2) + 0.3 \cdot \text{Cauchy}(-4, 1)$ where again $\theta = (\mu, \sigma) = (-4, 1.4)$. That is,

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} Q' \quad \text{where } Q' = 0.7 \cdot \mathcal{N}(-4, 1.4^2) + 0.3 \cdot \text{Cauchy}(-4, 1)$$

and $\mu = -4$, $\sigma = 1.4$ are the true parameters which have to be estimated. Numerical calculations show that, in fact,

$$Q' \in \mathcal{M}'_{(-4,1.4)} \tag{6.61}$$

even though Q' consists of a very strong – 30 per cent (!) – contamination with a Cauchy-distribution. On the one hand, this demonstrates that the credal sets $\mathcal{M}'_\theta$ are much larger than usual contamination neighborhoods with radius 0.02. On the other hand, this demonstrates that the use of such a (strongly contaminated) distribution Q' is not unreasonable because (6.61) implies that the Cauchy-distribution is extremely similar to the normal distribution – at least with respect to the probabilities of the above segments; and such probabilities are the only aspects of the normal distribution which are often taken into account e.g. by chi-square tests.

In the simulation study, our minimum distance estimator is compared with the classical estimators for μ and σ, the mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the empirical standard deviation

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Since the parameter set Θ is restricted, estimations which exceed the bounds of Θ are not reasonable. Therefore, the classical estimators are truncated by the bounds of Θ. For example, the result $\overline{x} = 5.7$ leads to the (truncated) estimation $\hat{\mu} = 5$.

Table 6.2 shows the (empirical) mean squared errors for both estimators in part 1 (“ideal situation”) and part 2 (“real situation”) of the simulation study. These values demonstrate that our minimum distance estimator behaves reasonable well in both cases while the classical estimators are quite perfect in the “ideal situation” but lead to unreliable results in real situations. This is also made visible by the boxplots shown in Figure 6.4. In

	MinDistance	Classical
ideal situation	0.10504	0.02818
real situation	0.10965	2.42263

Table 6.2: Empirical mean squared error calculated over the estimations obtained in 500 runs for each part of the simulation study in Model 2

particular, it can be seen that the classical estimation of the standard deviation σ breaks down in the “real situation” – only the bounded parameter set and the implemented truncation prevent it from exploding estimations. The contamination with a Cauchy distribution does not have any essential influence on our minimum distance estimator. This is not surprising at all because the minimum distance estimator does not really keep an eye on the normal distributions. It is only concerned with the probabilities of some segments and, with respect to these probabilities, normal distributions and Cauchy distributions are nearly the same.

6.6.3 Model 3: Linear regression

In order to demonstrate that the proposed minimum distance estimator can be applied in many different situations, Model 3 is concerned with linear regression where the error distribution is not symmetric and not even has mean 0. This flexibility of the estimator is due to its conceptual simplicity.[15]

In Model 3, we have two explanatory variables $z^{(1)} \in [0,1]$ and $z^{(2)} \in [0,1]$; the response variable is $y \in \mathbb{R}$. That is, the observations are

$$x_i \;=\; (y_i, z_i) \;=\; \big(y_i\,,\big(z_i^{(1)}, z_i^{(2)}\big)\big) \;\in\; \mathbb{R} \times [0,1]^2\,, \qquad i \in \{1,\dots,n\}$$

and we have

$$y_i \;=\; z_i^{(1)}\theta_1 + z_i^{(2)}\theta_2 + \varepsilon_i\,, \qquad i \in \{1,\dots,n\}$$

where

$$\theta \;=\; \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \;\in\; \Theta \;:=\; [-5,5] \times [-1,1]$$

is an unknown two-dimensional parameter and ε_i is an unobservable random error.

In the classical setup, it is most often assumed that the distribution of the errors ε_i is a normal distribution or, at least, that the distribution is symmetric around 0. Though the latter assumption is reasonable in many situations, this is certainly not always true. Positive errors may be more likely than negative errors (or vice versa) in many situations. So, the errors may, for example, be independent identically distributed according to the precise distribution S_0 which is a shifted log-normal distribution with Lebesgue-density

$$\frac{dS_0}{d\lambda} \;:\; \mathbb{R} \;\to\; [0,\infty)\,, \qquad x \;\mapsto\; \frac{2}{x\sqrt{2\pi}}\, e^{-2\left(\ln\left(x+e^{-0.25}\right)\right)^2}$$

[15] Skewed distributions are also considered e.g. in the theory of generalized linear models; cf. McCullagh and Nelder (1983).

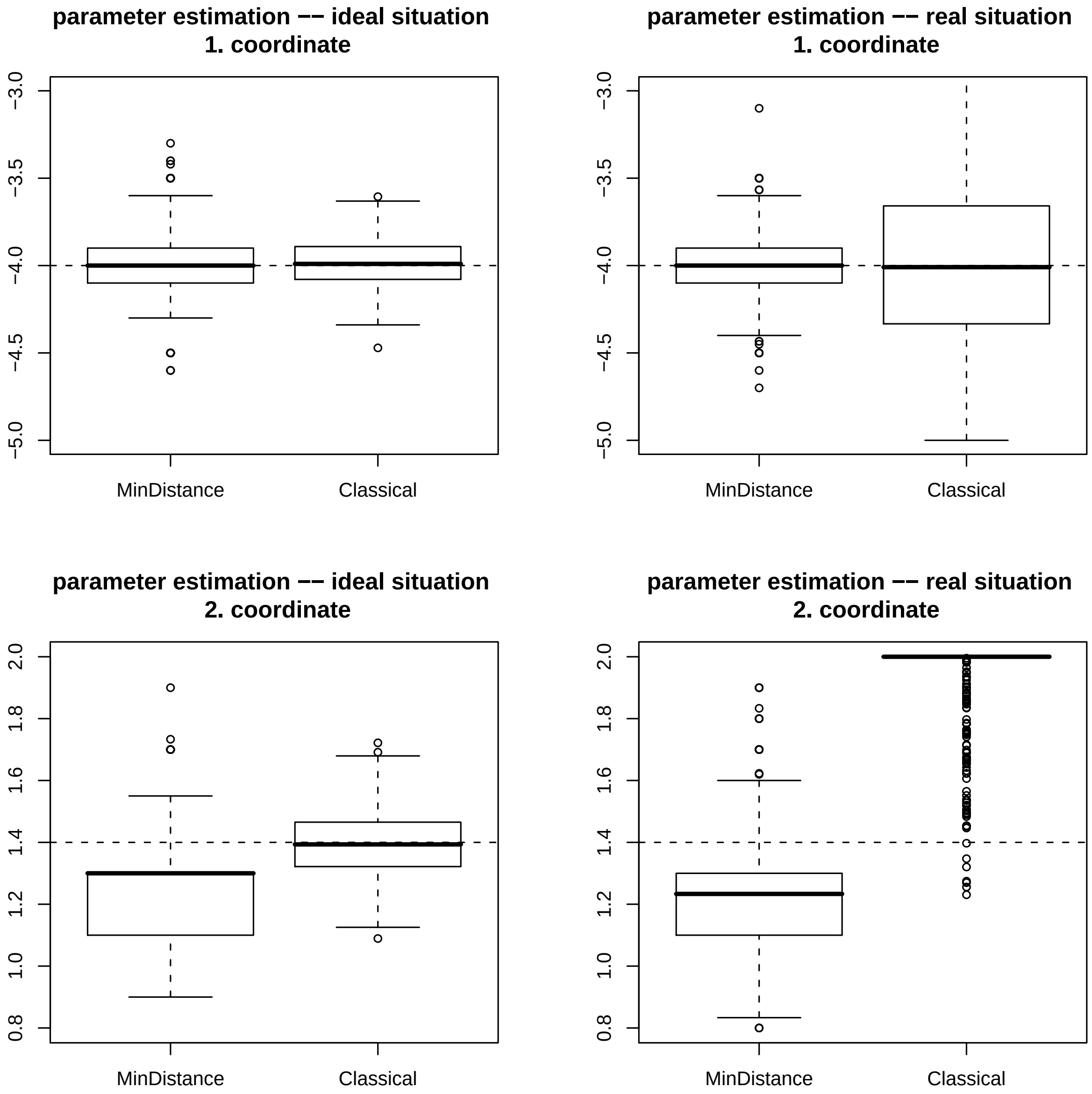

Figure 6.4: Boxplots of the estimations of both parameters obtained in 500 runs for each part of the simulation study in Model 2

That is, the log-normal distribution Log–N(μ,σ^2) with $\mu = 0$ and $\sigma = 0.5$ is shifted by $e^{\mu-\sigma^2}$ to the right so that the mode is equal to 0. The density of S_0 is pictured in Figure 6.6. Of course, assuming that the errors would precisely be distributed according to such an ideal distribution is as hazardous as assuming a normal distribution. Therefore, we go over to the upper coherent prevision

$$\overline{S} \;:\; \mathcal{L}_\infty(\mathbb{R},\mathfrak{B}) \;\to\; \mathbb{R}\,, \qquad h \;\mapsto\; \overline{S}[h]$$

which is based on the probabilities of some segments of the real line as follows: Put

$$h_0 \;=\; I_{(-\infty,-1]}\,, \quad h_1 \;=\; I_{(-1,-0.75]}\,, \quad \ldots\;, \quad h_j \;=\; I_{(-1+0.25(j-1)\,,\,-1+0.25j]}\,, \quad \ldots\;,$$

$$h_{17} \;=\; I_{(3,\infty)}$$

The credal set $\mathcal{N}$ of $\overline{S}$ consists of all probability charges S on $\mathbb{R}$ such that

$$S[h_0] \;\le\; (1-r)S_0[h_0] + r\,, \qquad S[h_{17}] \;\le\; (1-r)S_0[h_{17}] + r$$

and

$$(1-r)S_0[h_j] \;\le\; S[h_j] \;\le\; (1-r)S_0[h_j] + r \qquad \forall\, j \in \{1,\ldots,16\}$$

where

$$r \;=\; 0.05$$

Though the credal set $\mathcal{N}$ defined in this way is seemingly very similar to an ordinary contamination neighborhood used in robust statistics, $\mathcal{N}$ is much larger than such a neighborhood because, as in Model 2, the definition of $\mathcal{N}$ takes only the probabilities of some segments of the real line into account.

The parameter set Θ is again discretized:

$$\Theta_0 \;=\; \left\{ (\theta_1,\theta_2)\in\Theta \;\middle|\; \begin{array}{lcl} \theta_1 &=& -5 + 0.1\cdot k_1 - 0.05\,,\; k_1 \in \{1,\ldots,100\}\,, \\ \theta_2 &=& -1 + 0.1\cdot k_2 - 0.05\,,\; k_2 \in \{1,\ldots,20\} \end{array} \right\}$$

That is, $\theta_0 = (\theta_1,\theta_2)$ corresponds to the rectangle $(\theta_1 - 0.05, \theta_1 + 0.05] \times (\theta_2 - 0.05, \theta_2 + 0.05]$ with center $\theta_0 = (\theta_1,\theta_2)$.

The coherent upper prevision $\overline{S}$ would define an imprecise model for the distribution of (y,z) via

$$\overline{P}'_\theta[f_{\theta,j}] \;=\; \overline{S}[h_j]$$

where

$$f_{\theta,j}(y,z) \;=\; h_j\Big(y - z_i^{(1)}\theta_1 - z_i^{(2)}\theta_2\Big)$$

However, we do not assume stochastic explanatory variables – just as done in classical linear regression where the observed data $z_1,\ldots,z_n$ for the explanatory variable are used to build the design matrix. That is, the observed data $z_1,\ldots,z_n$ are not treated as data but determine the model. In order to adopt such a proceeding here, the imprecise model $(\overline{P}'_\theta)_{\theta\in\Theta}$ has to be defined in the following way:

For every $\theta_0 = (\theta_1,\theta_2) \in \Theta_0$ and every $j \in \{0,\ldots,17\}$

$$f_{\theta_0,j} \;:\; \mathbb{R}\times\{1,\ldots,n\} \;\to\; \mathbb{R}\,, \qquad (y,i) \;\mapsto\; h_j\Big(y - z_i^{(1)}\theta_1 - z_i^{(2)}\theta_2\Big)$$

	MinDistance	LeastSquares
empirical MSE	0.1472056	0.2055727

Table 6.3: Empirical mean squared error calculated over the estimations obtained in 100 runs in Model 3

Then, the credal set $\mathcal{M}'_{\theta_0}$ of $\overline{P}'_{\theta_0}$ consists of all probability charges P' on $\mathbb{R} \times \{1, \ldots, n\}$ such that

$$P'[f_{\theta_0,0}] \ \le \ \overline{S}[h_0] \, , \qquad P'[f_{\theta_0,17}] \ \le \ \overline{S}[h_{17}]$$

and

$$\underline{S}[h_j] \ \le \ P'[f_{\theta_0,j}] \ \le \ \overline{S}[h_j] \qquad \forall\, j \in \{1, \ldots, 16\}$$

where $\underline{S}$ is the coherent lower prevision which corresponds to $\overline{S}$.
In this way, the observed explanatory variables $z_1, \ldots, z_n$ determine the model and the according data in this model (where the explanatory variables are not assumed to be stochastic entities) are

$$(y_1, 1) \, , \ \ldots \ , \ (y_n, n)$$

and our minimum distance estimator can be applied on these data in the imprecise model $(\overline{P}'_{\theta_0})_{\theta_0 \in \Theta_0}$.

This has been done for sample size $n = 250$ in the simulation study. The errors stem from the shifted log-normal distribution S_0 . So, the error distribution is neither symmetric nor has mean zero. The true parameter is $\theta = (-4, 0.5)$. Our minimum distance estimator is compared to the classical least-square estimator. Due to rather high computational costs, only 100 runs have been made for this model.

The boxplots in Figure 6.5 demonstrate that the applied error distribution leads to a considerable bias of the least-squares estimator whereas the minimum distance estimator is not biased. However, the variance of the minimum distance estimator is still quite large. This is because the imprecise model is only based on the probabilities of some very few segments of the real line so that the imprecise model is rather nonparametric than parametric. In order to obtain better results of our minimum distance estimator, the sample size n has to be increased in this nonparametric situation. Of course, increasing the sample size does not improve the results of the least-squares estimator. Table 6.3 shows that the empirical MSE of the minimum distance estimator is considerably smaller than the one of the least-squares estimator.

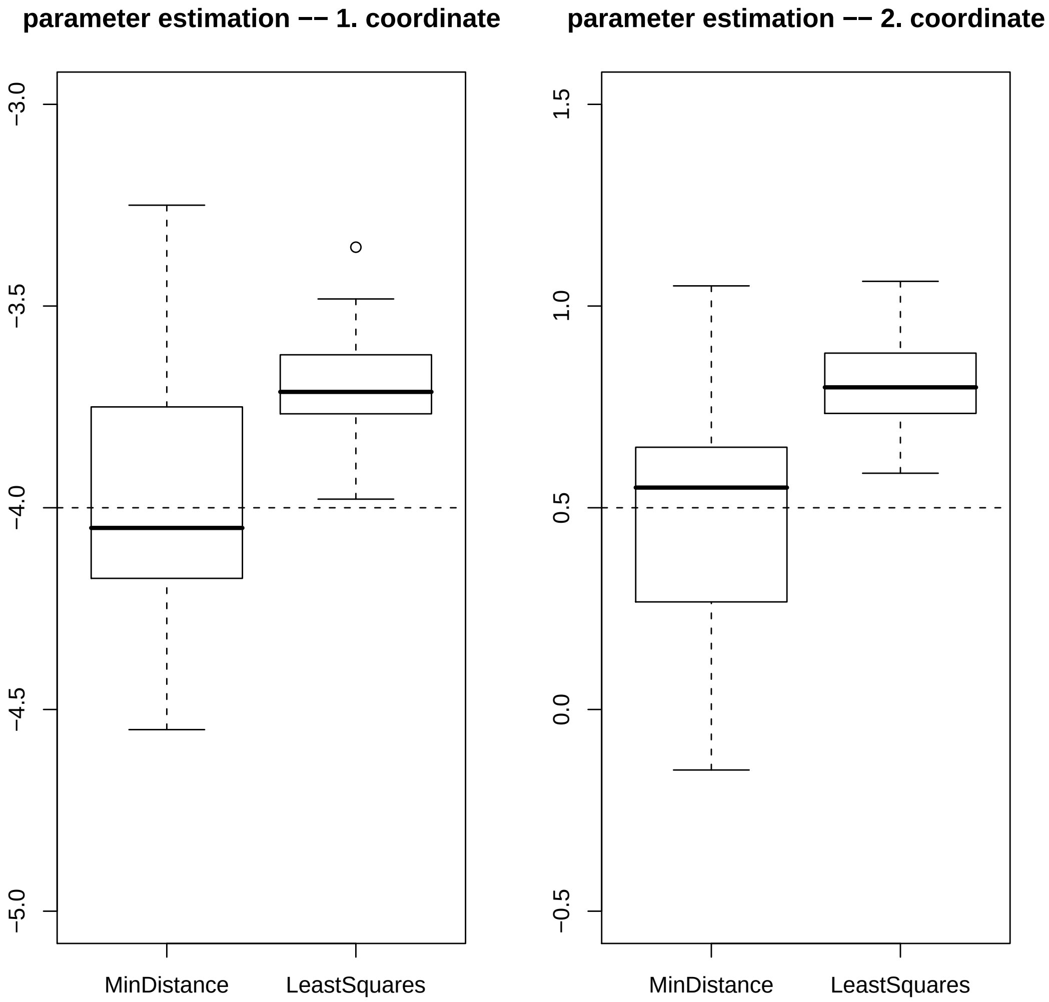

Figure 6.5: Boxplots of the estimations of both parameters obtained in 100 runs in Model 3

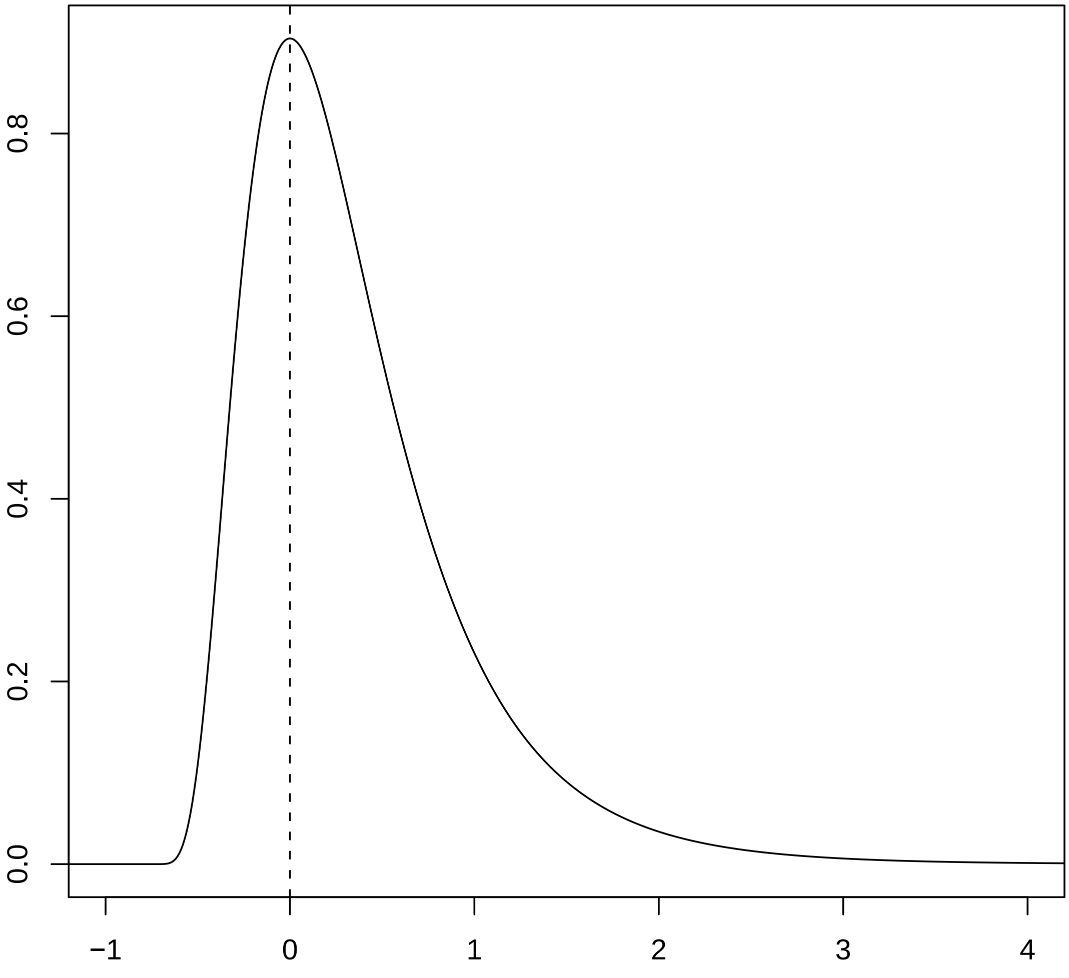

Figure 6.6: Densitiy $x \mapsto \frac{2}{x\sqrt{2\pi}} e^{-2\left(\ln\left(x+e^{-0.25}\right)\right)^2}$ of the error distribution which is a shifted log-normal distribution

Chapter 7

Conclusion and outlook

The present book is concerned with data-based decision problems under imprecise probabilities. Using imprecise probabilities in decision problems is advisable in many situations because the arising uncertainties are often much too complex to be adequately modeled by classical precise probabilities. On the one hand, this leads to more realistic models and more reliable results. On the other hand, this increases the mathematical input for solving decision problems. This is true the more so as going over to imprecise probabilities in a frequentist setting makes it necessary to consider decision problems with an explicit data-based formulation as explained in Vidakovic (2000) and Augustin (2003). Nevertheless, this topic has hardly been investigated before even though many recent publications are concerned with data-free decision problems under imprecise probabilities. Therefore, the present book cannot provide a final disquisition on this new topic but may serve as a sound starting point for further research.

Accordingly, the book starts with some basic groundwork: First of all, topological properties of different concepts of imprecise probabilities are investigated and compared to each other. This is most fundamental with respect to their application in decision theory because, there, using minimax theorems which rely on topology is essential. As a result, the concept of coherent upper previsions turns out to have such topological properties which are more desirable than the ones of F-probabilities. These investigations also lead to the first explicit treatment of F-probabilities on Polish spaces and compact Hausdorff spaces. Since F-probabilities are consciously developed in the style of classical measure theory, using these classical setups of topological measure theory is suggesting. As an intermediate step between coherent upper previsions and F-probabilities, upper expectations are considered which have original been defined by Buja (1984) in robust statistics but have not been considered within the theory of imprecise probabilities before. There is only one difference between upper expectations and coherent upper previsions: While upper expectations rely on σ-additivity the concept of coherent upper previsions dispenses with σ-additivity. However, it is shown in the present book that every coherent upper prevision can be represented by a (canonical) upper expectation on a compact Hausdorff space. This offers an interesting tool for future research on coherent upper previsions: Dispensing with σ-additivity makes it hard or even impossible to carry over concepts of classical probability theory which rely on σ-additivity. Now, this representation provides a possibility to come around this problem in a canonical way. This may, in particular, contribute to investigations on conditional imprecise probabilities – a topic in which searching for suitable definitions is still a matter of research. Since coherent upper previsions are mathematically equivalent to risk measures in mathematical finance, these results can

also be used to represent risk measures by sets of σ-additive probability measures. Such representations have recently received considerable attention in mathematical finance, cf. for example Delbaen (2002), Föllmer and Schied (2004) and Krätschmer (2005).
After this comparison of different concepts of imprecise probabilities, an extended decision theoretic framework under imprecise probabilities is developed. This framework contains several decision theoretic tools which are essential for many results obtained in the present book. Therefore, it seems to be most likely that they can profitably be used in the theory of imprecise probability further on. This is also indicated by the fact that these tools are mainly based on concepts developed by L. Le Cam in order to deal with large models – and imprecise probabilities, in fact, lead to large models. For example, as a welcomed byproduct, the present book suggests a definition of sufficiency in case of imprecise probabilities for the first time. Though elaborated applications would have been out of the scope of the book and are a matter of future research, it is demonstrated how sufficiency can be used in order to deal with parametrically generated imprecise models. In this way, the proposed notion of sufficiency could also be applied in the popular Imprecise Dirichlet Model.[1]

Since Buja (1984) is concerned with a very similar setup – namely data-based decision theory where uncertainties are modeled by upper expectations, this article is revised within the only recently available theory of imprecise probabilities. It is shown that, due to an erroneous statement, the results of Buja's article are only assured in case of an undesirable extra assumption. It is proven that such an extra assumption can be avoided by going over to the setup used in the present book. Here, the desirable topological properties of coherent upper previsions pays off. This leads to a generalization of results within the Huber-Strassen theory: A necessary and sufficient condition for the existence of least favorable models is given. This offers a general tool which makes it possible to reduce the computational effort in data-based decision theory under imprecision. However, further research has to be done for using it in concrete problems: As in Huber and Strassen (1973), this result is only concerned with the existence of least favorable models but an algorithm for calculating least favorable models has not yet been developed. After Huber and Strassen (1973), a lot of work was done to construct least favorable pairs in hypothesis testing (confer e.g. Rieder (1977), Österreicher (1978), Hafner (1992), Augustin (1998)). In the more general case of the present book, this is a matter of further research.

Nevertheless, results obtained by these investigations are applied afterwards in order to justify the use of the method of natural extension, which is fundamental within the theory of imprecise probabilities, in data-based decision problems. It is shown by means of the theory of vector lattices that applying the method of natural extension in decision problems does not affect the optimality of decisions. However, it is also shown that, in general, the method of natural extension suffers from a severe instability: Arbitrarily small changes in coherent upper previsions can have arbitrarily large effects on their natural extension and, therefore, applying natural extensions may lead to meaningless results. This is unfortunate the more so as imprecise probabilities are intended to prevent from such unreliable results. However, not all is lost since it can be guaranteed in many situation that small changes in the coherent upper previsions have small effects on the natural extension. However, these results are not fully satisfactory; hopefully, these initial investigations serve as a starting point for future research into this direction.

[1] For the Imprecise Dirichlet Model, confer e.g. Walley (1996), Bernard (2005) and the forthcoming special issue of the *International Journal of Approximate Reasoning* on the Imprecise Dirichlet Model.

The book closes with parameter estimation as an application in statistics. This is a topic which has hardly been considered explicitly within the theory of coherent upper previsions so far. Since we are not yet able to calculate optimal estimators within this setup, a minimum distance estimator is developed which is proven to have some good properties. An algorithm for calculating the estimator is given which is based on linear programming and the applicability of the estimator is verified by a simulation study. In particular, the simulation study shows that the proposed estimator can even be used for large sample sizes and may, in fact, lead to good results in realistic situations. This meets objections that imprecise probabilities could not be used for practical purposes. Due to the present state of research, this work cannot be restricted to the sole investigation of the proposed estimator but also has to develop some fundamentals of (frequentist) estimating under coherent upper previsions at first. This is necessary the more so as the minimum distance estimator is associated with the empirical process (which needs a somewhat more elaborated setting) and is justified by asymptotic arguments (but an elaborated asymptotic theory of imprecise probabilities is still missing). In doing so, the work also provides a base for future research about estimating under imprecise probabilities. In particular, it would be desirable to develop alternative estimators so that the proposed minimum distance estimator can be compared to other estimators under imprecise probabilities. Furthermore, the simulation study presented in this book is only intended to demonstrate the applicability of the proposed minimum distance estimator but it would have been out of the scope to investigate more advanced applications in involved statistical models. For example, robust statistics is already able to deal with involved semiparametric regression models such as the Cox model in survival analysis (cf. e.g. Sasieni (1993), Bednarski and Nowak (2003) and Hable et al. (2008)) but Model 3 in the simulation study of the present book is only concerned with two-dimensional linear regression. Nevertheless, the simulation study indicates that the flexibility of the proposed minimum distance estimator enables future applications in more advanced models. In order to encourage this, the estimator has been programmed in R and has already been made publicly available as (open source) R package “imprProbEst”; cf. Hable (2008a).

Chapter 8

Appendix

8.1 Vector lattices

This section presents the basics of the theory of vector lattices and is mainly based on (Bhaskara Rao and Bhaskara Rao, 1983, Section 1.5). Important notions related to vector lattices are bands and L-spaces.

Definition 8.1 *Let V be a vector space which is endowed with a partial ordering $\leq$ such that*

$$\begin{array}{lll} x, y \in V, \quad x \leq y & \Rightarrow & x + z \leq y + z \quad \forall\, z \in V \\ x, y \in V, \quad x \leq y & \Rightarrow & cx \leq cy \quad \forall\, c \in [0, \infty) \end{array}$$

Then V is called an ordered vector space.

Definition 8.2 *Let V be an ordered vector space. For $x, y \in V$, an element z with $x \leq z$, $y \leq z$ and the property*

$$\hat{z} \in V, \quad x \leq \hat{z}, \ y \leq \hat{z} \qquad \Rightarrow \qquad z \leq \hat{z}$$

is called supremum of x and y. It is denoted by $x \vee y$.
If the supremum of x and y exists, it is unique. An infimum of x and y is analoguously defined. It is denoted by $x \wedge y$.
$w \in V$ is called majorant of a subset $S \subset V$ if $s \leq w \quad \forall\, s \in S$. S is called majorised or order bounded then.
A majorant $w \in V$ of $S \subset V$ is called supremum of S if

$$\hat{w} \in V, \quad s \leq \hat{w} \ \forall\, s \in S \qquad \Rightarrow \qquad w \leq \hat{w}$$

Definition 8.3 *A vector space V is called vector lattice if it is an ordered vector space so that, $x \vee y$ and $x \wedge y$ exist for every $x, y \in V$.*

Let V be a vector lattice. For $x \in V$, $x^+ := x \vee 0$ is called positive part of x, $x^- := -(x \wedge 0)$ is called negative part of x and $|x| := x^+ + x^-$ is called modulus of x. The following assertions are valid for every $x \in V$:

$$x = x^+ - x^-$$

$$x^+ \wedge x^- = 0$$

$$|x| = x^+ \vee x^-$$

$x, y \in V$ are called orthogonal, if $|x| \wedge |y| = 0$. Orthogonality of x and y is denoted by $x \perp y$.

Definition 8.4 *A sublattice W of a vector lattice V is a vector subspace of V so that $x \vee y,\, x \wedge y \in W \;\; \forall\, x, y \in W$.*

Definition 8.5 *A subset $B \subset V$ of a vector lattice V is called band if*

(i) *B is a sublattice of V,*

(ii) *$x \in B,\, 0 \leq |y| \leq |x| \;\; \Rightarrow \;\; y \in B$*

(iii) *each nonempty subset of B that has a majorant in V also has a supremum, which belongs to B.*

Definition 8.6 *A vector lattice V is called order complete or Dedekind complete or boundedly complete if every majorised subset of V has a supremum in V.*

Remark 8.7 *The intersection of arbitrarily many bands is again a band. Every subset $S \subset V$ of a Dedekind complete vector lattice V has a smallest band containing S.*

For a subset $S \subset V$ of a vector lattice V, define

$$S^\perp = \big\{ x \in V \;\big|\; x \perp s \;\; \forall\, s \in S \big\}$$

Proposition 8.8 *Let $S \subset V$ be a subset of a Dedekind complete vector lattice V. Then, $S^\perp$ is a band.*

Corollary 8.9 *Let $S \subset V$ be a subset of a Dedekind complete vector lattice V. Then, $(S^\perp)^\perp$ is the smallest band containing S.*

Theorem 8.10 (Riesz Decomposition Theorem) *Let B be a band in a Dedekind complete vector lattice V. Then,*

$$V \;=\; B \oplus B^\perp$$

i.e.: for every element $x \in V$, there is a unique $x' \in B$ and a unique $x'' \in B^\perp$ so that

$$x \;=\; x' + x'' \tag{8.1}$$

The maps

$$\pi_B : \;\; V \to B, \quad x \mapsto x' \qquad \textit{and} \qquad \pi_{B^\perp} : \;\; V \to B^\perp, \quad x \mapsto x''$$

have the following properties:

- *π_B and $\pi_{B^\perp}$ are linear.*
- $\pi_B(x') \;=\; x' \quad \forall\, x' \in B\,, \qquad \pi_{B^\perp}(x'') \;=\; x'' \quad \forall\, x'' \in B^\perp$
- $x \,\geq\, 0 \qquad \Rightarrow \qquad \pi_B(x) \,\geq\, 0\,, \quad \pi_{B^\perp}(x) \,\geq\, 0$
- $x \;=\; \pi_B(x) \,+\, \pi_{B^\perp}(x) \qquad \forall\, x \in V$

If $x \geq 0$, then $x' = \pi_B(x)$ is the supremum of $\left\{ b \in B \;\middle|\; 0 \leq b \leq x \right\}$. Therefore,

$$\pi_B(x) = \bigvee_{s \in S} \big(x \wedge |s| \big)$$

Generally, $\pi_B(x) = \pi_B(x^+) - \pi_B(x^-)$ for $x \in V$.

Definition 8.11 *A vector lattice V provided with a norm $\|\cdot\|$ is said to be a normed vector lattice if the norm is compatible with the modulus $|\cdot|$, i.e.:*

$$x, y \in V,\ |x| \leq |y| \qquad \Rightarrow \qquad \|x\| \leq \|y\|$$

If, in addition, V is a Banach space, then it is also called a Banach lattice.

The norm of a normed vector lattice V induces the norm-topology on V.

Proposition 8.12 *Let V be a normed vector lattice. Then, the maps*

$$(x, y) \mapsto x \vee y, \qquad (x, y) \mapsto x \wedge y$$

are norm-continuous. For every subset $S \subset V$, $S^\perp$ is norm-closed. Every band $B \subset V$ is norm-closed.

Definition 8.13 *An L-space V is a Banach lattice where*

$$x, y \in V, \quad x \geq 0,\ y \geq 0 \qquad \Rightarrow \qquad \|x + y\| = \|x\| + \|y\|$$

Theorem 8.14 *Every L-space is Dedecind complete.*[1]

Remark 8.15 *Every band $B \subset V$ in an L-space V is itself an L-space.*

Definition 8.16 *An M-space V is a Banach lattice where*

$$x, y \in V, \quad x \geq 0,\ y \geq 0 \qquad \Rightarrow \qquad \|x \vee y\| = \|x\| \vee \|y\|$$

Definition 8.17 *Let V and W be vector lattices. A map $\varphi : V \to W$ is called* vector lattice isomorphism *or* isomorphism of vector lattices *if it is*

- *linear*
- *bijective*
- $\varphi(x \wedge y) = \varphi(x) \wedge \varphi(y)\,, \quad \varphi(x \vee y) = \varphi(x) \vee \varphi(y)\,, \qquad \forall\, x, y \in V$
- $\varphi^{-1}(w \wedge z) = \varphi^{-1}(w) \wedge \varphi^{-1}(z)\,, \quad \varphi^{-1}(w \vee z) = \varphi^{-1}(w) \vee \varphi^{-1}(z)\,,$ $\forall\, w, z \in W$

Definition 8.18 *Let V and W be Banach lattices. A map $\varphi : V \to W$ is called* Banach lattice isomorphism *or* isomorphism of Banach lattices *if it is*

- *a vector lattice isomorphism and*
- *isometric:* $\|\varphi(x)\| = \|x\| \quad \forall\, x \in V$

[1] Confer e.g. (Schaefer, 1974, Proposition 8.3(ii))

Definition 8.19 *Let V and W be Banach lattices. A Banach lattice isomorphism $\varphi : V \to W$ is called*

a) L-space isomorphism *or* isomorphism of L-spaces *if V and W are L-spaces.*

b) M-space isomorphism *or* isomorphism of M-spaces *if V and W are M-spaces.*

Proposition 8.20 [2] *Let V and W be vector lattices. A map $\varphi : V \to W$ is a vector lattice isomorphism if and only if it is*

- *linear*
- *bijective*
- $x \geq 0 \quad \Leftrightarrow \quad \varphi(x) \geq 0$

Proposition 8.21 *Let V and W be L-spaces. A map $\varphi : V \to W$ is an L-space isomorphism if and only if it is*

- *a vector lattice isomorphism and*
- *normalized:* $\|\varphi(x)\| = \|x\| \quad \forall\, x \geq 0\,,\ x \in V$

Proof: Let φ be a normalized vector lattice isomorphism. Then, for every $x \in V$,

$$\begin{aligned} \|\varphi(x)\| &= \big\|\, |\varphi(x)|\, \big\| = \big\|\big(\varphi(x)\big)^+ + \big(\varphi(x)\big)^-\big\| = \\ &= \big\|\big(\varphi(x)\big)^+\big\| + \big\|\big(\varphi(x)\big)^-\big\| = \|\varphi(x^+)\| + \|\varphi(x^-)\| = \\ &= \|\,x^+\| + \|\,x^-\| = \|\,x^+ + x^-\| = \big\|\,|x|\,\big\| = \|x\| \end{aligned}$$

because in a normed vector lattice is $\big\|\,|x|\,\big\| = \|x\|$ (cf. (Schaefer, 1974, p. 81)), i.e. φ is isometric.

The converse statement is trivial. □

Proposition 8.22

a) *The dual space of an L-space is an M-space. The dual space of an M-space is an L-space.*

b) *Let $\varphi : V \to W$ be an M-space isomorphism. Then, the adjoint of φ*

$$\varphi^* \; : \; W^* \;\to\; V^*$$

is an L-space isomorphism.

For part a), confer (Schaefer, 1974, Proposition 9.1); part b) is an easy corollary then.

[2] cf. (Constantinescu et al., 1998, Proposition 1.5.6 (c))

8.2 Weak topologies

This section introduces (a special case of) weak topologies following the lines of (Dunford and Schwartz, 1958, §V.3). Examples of such topologies which appear in this book are the weak*-topology on $\mathrm{ba}(\Omega, \mathcal{A})$ and the weak topology of probability measures.

Let Ω be a set and $\mathcal{L}_\infty(\Omega)$ be the Banach space of all bounded functions $f : \Omega \to \mathbb{R}$ with norm $\|f\| = \sup_{\omega\in\Omega} |f(\omega)|$. Let $\Gamma \subset \mathcal{L}_\infty(\Omega)$ be a linear subspace. Furthermore, let M be a linear subspace of the dual space Γ^*. So, every $\mu \in M$ is a continuous, linear funktional

$$\mu : \ \Gamma \ \to \ \mathbb{R}, \qquad f \ \mapsto \ \mu[f]$$

For every $f \in \Gamma$, put

$$\Lambda_f : \ M \ \to \ \mathbb{R}, \qquad \mu \ \mapsto \ \Lambda_f(\mu) \ = \ \mu[f]$$

which is a continuous (in the norm-topology), linear functional on M.

Definition 8.23 *The* Γ - topology on M *is the weakest topology on* M *such that every* Λ_f *is continuous (in this topology) for every* $f \in \Gamma$.
That is, the Γ - topology on M *is the weakest topology on* M *such that the sets*

$$\Lambda_f^{-1}(B), \quad f \in \Gamma\,, \ \ B \subset \mathbb{R} \textit{ open}$$

are open.

Now, M has two different topologies, namely the norm-topology and the (weaker) Γ-topology. To make clear what topology is used, topological terms such as compact, open, closure etc. usually are denoted by norm-compact, Γ-open, norm-closure etc.

According to (Dunford and Schwartz, 1958, Definition V.3.2 and Lemma V.3.8), the sets

$$N(\mu, F, \varepsilon) \ = \ \big\{\nu \;\big|\; |\mu(f) - \nu(f)| \ < \varepsilon\,, \ f \in F\big\}$$

where

$$\mu \in M\,, \quad F \text{ is a finite subset of } \Gamma\,, \quad \varepsilon > 0$$

form a base of the Γ - topology

The next theorem sumarizes some common properties of the Γ - topology according to (Dunford and Schwartz, 1958, p. 420f).

Theorem 8.24

- ***a)*** *M is a locally convex linear topological (Hausdorff) space in its Γ - topology.*
- ***b)*** *A net $(\mu_\beta)_{\beta\in B}$ converges to μ in the Γ - topology if and only if* $\lim_\beta \mu_\beta[f] = \mu[f]$ *for every $f \in \Gamma$.*
- ***c)*** *The linear functionals on M which are Γ-continuous are precisely the functionals $\Lambda_f : \mu \mapsto \mu[f]$, $f \in \Gamma$.*

Lemma 8.25 characterizes the subspace topology on $M_0 \subset M$ generated by the Γ-topology on M.

Lemma 8.25 *Let M_0 be a linear subspace of $\mathcal{L}_\infty(\Omega)$ such that $M_0 \subset M$. Then, the subspace topology on $M_0 \subset M$ generated by the Γ-topology on M is equal to the Γ-topology on M_0.*

Proof: Let $(\mu_\beta)_{\beta \in B}$ be a net in M_0 and $\mu \in M_0$. Then,

$$\begin{aligned} & \mu_\beta \underset{\beta}{\longrightarrow} \mu \in M_0 \quad \text{in the subspace topology on } M_0 \subset M \\ \Leftrightarrow \quad & \mu_\beta \underset{\beta}{\longrightarrow} \mu \in M_0 \quad \text{in the } \Gamma\text{-topology on } M \\ \Leftrightarrow \quad & \mu_\beta[f] \underset{\beta}{\longrightarrow} \mu[f] \quad \forall f \in \Gamma \\ \Leftrightarrow \quad & \mu_\beta \underset{\beta}{\longrightarrow} \mu \in M_0 \quad \text{in the } \Gamma\text{-topology on } M_0 \end{aligned}$$

□

Just for a moment, put $M = \mathrm{ba}(\Omega, \mathcal{A})$ and $\Gamma = \mathcal{L}_\infty(\Omega, \mathcal{A})$. Then, the lower envelope theorem (Walley, 1991, Theorem 2.6.3 (b)) states that

$$\overline{P}[f] = \sup_{\mu \in \mathcal{V}} \mu[f], \qquad \forall f \in \mathcal{L}_\infty(\Omega, \mathcal{A})$$

defines a coherent upper prevision for $\mathcal{V} \subset \mathrm{ba}(\Omega, \mathcal{A})$ and the weak*-compactness theorem (Walley, 1991, Theorem 3.6.1) implies that the credal set of $\overline{P}$ is equal to the convex $\mathcal{L}_\infty(\Omega, \mathcal{A})$ - closure of $\mathcal{V}$.

The following theorem is a generalization of this result for arbitrary linear subspaces $\Gamma \subset \mathcal{L}_\infty(\Omega)$ and $M \subset \Gamma^*$. – It is needed for upper expectations and F-probabilities where weak topologies on $\mathrm{ca}_1^+(\Omega, \mathcal{A})$ are considered.

Theorem 8.26 *Let $\mathcal{V} \subset M$ be any subset of M. Put*

$$\overline{P}[f] = \sup_{\mu \in \mathcal{V}} \mu[f], \qquad \forall f \in \Gamma$$

Then, the convex Γ - closure of $\mathcal{V}$ is

$$c\ell_\Gamma co\big(\mathcal{V}\big) = \big\{\mu \in M \;\big|\; \mu[f] \le \overline{P}[f] \;\; \forall f \in \Gamma\big\} := \mathcal{M}$$

Proof: If $\overline{P}[f] < \infty$, put $I_f = \big(-\infty\,,\, \overline{P}[f]\big]$; if $\overline{P}[f] = \infty$, put $I_f = \mathbb{R}$.

Since every $\Lambda_f : \mu \mapsto \mu[f]$ is Γ-continuous and each interval I_f is closed in $\mathbb{R}$,

$$\mathcal{M} = \bigcap_{f \in \Gamma} \underbrace{\Lambda_f^{-1}(I_f)}_{\Gamma\text{-closed}}$$

is Γ-closed. That is, $\mathcal{M}$ is a Γ-closed convex set which contains $\mathcal{V}$ and, therefore, $c\ell_\Gamma \mathrm{co}\big(\mathcal{V}\big) \subset \mathcal{M}$.

Conversely, take any $\mu \in M$ such that $\mu \notin c\ell_\Gamma \mathrm{co}\big(\mathcal{V}\big)$. M is a locally convex linear topological space in the Γ-topology (Theorem 8.24), $c\ell_\Gamma \mathrm{co}\big(\mathcal{V}\big)$ is a Γ-closed convex subset and $\{\mu\}$ is a Γ-compact convex subset such that $\{\mu\} \cap c\ell_\Gamma \mathrm{co}\big(\mathcal{V}\big) = \emptyset$. According to (Dunford and Schwartz, 1958, Theorem V.2.10), there is a Γ-continuous linear functional $T : M \to \mathbb{R}$ such that

$$\sup_{\nu \in c\ell_\Gamma \mathrm{co}(\mathcal{V})} T(\nu) < T(\mu) \tag{8.2}$$

Next, Theorem 8.24 c) implies existence of some $f \in \Gamma$ such that $\Lambda_f = T$. Hence,

$$\overline{P}[f] \leq \sup_{\nu \in c\ell_\Gamma \mathrm{co}\,(\mathcal{V})} \nu[f] \overset{(8.2)}{<} \mu[f]$$

That is, $\mu \notin \mathcal{M}$. Hence, $\mathcal{M} \setminus c\ell_\Gamma \mathrm{co}\big(\mathcal{V}\big) = \emptyset$. □

8.3 Some technical lemmas

Lemma 8.27 *Let Ξ be a Polish space with Borel-σ-algebra $\mathfrak{B}$ and let P be a probability measure on $(\Xi, \mathfrak{B})$. Then, for every $f \in \mathcal{L}_\infty(\Xi, \mathfrak{B})$, there is a sequence of upper semicontinuous functions $(f_n)_{n\in\mathbb{N}} \subset \mathcal{L}_\infty(\Xi, \mathfrak{B})$ such that*

$$f_1 \leq f_2 \leq f_3 \leq \ldots \leq f \qquad \text{and} \qquad P[f_n] \nearrow P[f]$$

Proof: According to Lusin's Theorem (Bauer, 2001, Theorem 26.7), there is a sequence of compact subsets $(K_n)_{n\in\mathbb{N}}$ of Ξ such that

$$P\big(\Xi \setminus K_n\big) \leq \frac{1}{n} \qquad \forall\, n \in \mathbb{N}$$

and the restriction of f on K_n is continuous for every $n \in \mathbb{N}$. It is easy to see that each map

$$\inf f + \big(f - \inf f\big) I_{K_n}$$

is upper semicontinuous. Since the maximum of a finite number of upper semicontinuous functions is again upper semicontinuous, the functions

$$f_n := \max_{i \in \{1,\ldots,n\}} \Big(\inf f + \big(f - \inf f\big) I_{K_n}\Big), \qquad n \in \mathbb{N}$$

are upper semicontinuous. Furthermore,

$$f_1 \leq f_2 \leq f_3 \leq \ldots \leq f$$

and

$$0 \leq \limsup_n P[f_n] - P[f] \leq \limsup_n P\big(\Xi \setminus K_n\big) \cdot 2\|f\| \leq \limsup_n \frac{2}{n}\|f\| = 0$$

□

Lemma 8.28 *Let $\mathcal{B}$ be an algebra on a set $\mathcal{Y}$, let $\mathcal{D}$ be an algebra on a set $\mathcal{Z}$ and $\Psi : \mathcal{B} \to \mathcal{D}$ an algebra homomorphism. Put*

$$\zeta(I_B) = I_{\Psi(B)} \qquad \textit{for every } B \in \mathcal{B}$$

$$\zeta\Big(\sum_{j=1}^m b_j I_{B_j}\Big) = \sum_{j=1}^m b_j I_{\Psi(B_j)} \qquad \textit{for every simple function on } (\mathcal{Y}, \mathcal{B})$$

and

$$\zeta(g) = \lim_{n\to\infty} \zeta(s_n) \qquad \textit{for every } g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$$

where each s_n is a simple function and $\|s_n - g\| \to 0$.
Then,

$$\zeta : \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B}) \to \mathcal{L}_\infty(\mathcal{Z}, \mathcal{D}), \qquad g \mapsto \zeta(g)$$

is a well defined map which is

- *linear*
- *positive:* $\xi(h) \,\geq\, 0 \qquad \forall\, h \geq 0$
- *normalized:* $\xi(I_{\Xi}) \,=\, I_{\Omega}$

Proof: Note that $\Psi(\emptyset) = \emptyset$ and $\Psi(\mathcal{Y}) = \mathcal{Z}$.

Then, Ψ defines a map

$$\mathcal{Z} \times \mathcal{B} \,\rightarrow\, \mathbb{R}\,, \qquad (z,B) \,\mapsto\, I_{\Psi(B)}(z) \tag{8.3}$$

such that

- the function $z \,\mapsto\, I_{\Psi(B)}(z)$ is an element of $\mathcal{L}_\infty(\mathcal{Z},\mathcal{D})$ for every $B \in \mathcal{B}$.
- $\hat{\zeta}_z \,:\, B \,\mapsto\, I_{\Psi(B)}(z)$ is a probability charge on $\mathcal{B}$ for every $z \in \mathcal{Z}$.[3]

The first statement is obvious. The second statement is an easy consequence of the properties of Ψ:

$$\hat{\zeta}_z(\emptyset) \,=\, I_{\Psi(\emptyset)}(z) \,=\, 0\,, \qquad \hat{\zeta}_z(B) \,\geq\, 0 \quad \forall\, B \in \mathcal{B}$$

and for every $B_1,\, B_2 \,\in\, \mathcal{B}$ such that $B_1 \cap B_2 \,=\, \emptyset$,

$$\begin{aligned}\hat{\zeta}_z(B_1 \cup B_2) \;&=\; I_{\Psi(B_1 \cup B_2)}(z) \;=\; I_{\Psi(B_1)\cup\Psi(B_2)}(z) \;=\\ &=\; I_{\Psi(B_1)}(z) + I_{\Psi(B_2)}(z) \;=\; \hat{\zeta}_z(B_1) + \hat{\zeta}_z(B_2)\end{aligned}$$

Since $\hat{\zeta}_z$ is a probability charge for every $z \in \mathcal{Z}$, we can define the map

$$\hat{\zeta}(g) \,:\, \mathcal{Z} \,\rightarrow\, \mathbb{R}\,, \qquad z \,\mapsto\, \hat{\zeta}(g)(z) \tag{8.4}$$

for every $g \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ where

$$\hat{\zeta}(g)(z) \,=\, \int g(y)\, \hat{\zeta}_z(dy) \qquad \forall\, z \in \mathcal{Z}\,, \qquad g \,\in\, \mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \tag{8.5}$$

Then, for every simple function $\sum_{j=1}^m b_j I_{B_j}$ on $(\mathcal{Y},\mathcal{B})$,

$$\hat{\zeta}\Big(\sum_{j=1}^m b_j I_{B_j}\Big) \,=\, \sum_{j=1}^m b_j \int I_{\Psi(B_j)}(z)\, \hat{\zeta}_z(dy) \,=\, \sum_{j=1}^m b_j I_{\Psi(B_j)} \tag{8.6}$$

is a simple function on $(\mathcal{Z},\mathcal{D})$. Now, take any $g \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$. Then, there is a sequence of simple functions s_n such that $\|s_n - g\| \rightarrow 0$ and

$$\sup_z \big|\, \hat{\zeta}(s_n)(z) - \hat{\zeta}(g)(z)\,\big| \,\leq\, \sup_z \int \|s_n - g\|\, \hat{\zeta}_z(dy) \,=\, \|s_n - g\| \,\underset{n}{\longrightarrow}\, 0 \tag{8.7}$$

This implies that $\hat{\zeta}(g)$ is bounded and $\hat{\zeta}(g) \,\in\, \mathcal{L}_\infty(\mathcal{Z},\mathcal{D})$.

Hence,

$$\hat{\zeta} \,:\, \mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \,\rightarrow\, \mathcal{L}_\infty(\mathcal{Z},\mathcal{D})\,, \qquad g \,\mapsto\, \hat{\zeta}(g)$$

[3] The map defined by (8.3) essentially behaves like an ordinary Markov kernel. However, it is not exactly a Markov kernel because $\mathcal{B}$ is, in general, not a σ-algebra but an algebra.

is a well defined map which is obviously linear, positive and normalized.

Finally, it follows from (8.5) and (8.7) that

$$\zeta(g) \;=\; \hat{\zeta}(g) \qquad \forall\, g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$$

□

For ease of reference, the following lemma is proven:

Lemma 8.29 *Let X be a linear topological space and $\varphi : X \to \mathbb{R}$ a linear, continuous function. Then, for any subset $A \subset X$,*

$$\sup\big\{\varphi(x) \,\big|\, x \in A\big\} \;=\; \sup\big\{\varphi(x) \,\big|\, x \in c\ell\mathrm{co}(A)\big\}$$

where $c\ell\mathrm{co}(A)$ denotes the closed convex hull of A in X .

Proof: According to (Dunford and Schwartz, 1958, Theorem V.2.1 (a)), the closed convex hull $c\ell\mathrm{co}(A)$ is equal to the closure of the convex hull $\mathrm{co}(A)$ of A, i.e.

$$c\ell\mathrm{co}(A) \;=\; c\ell\big(\mathrm{co}(A)\big)$$

Hence, it suffices to show

$$\sup\big\{\varphi(x) \,\big|\, x \in A\big\} \;\geq\; \sup\big\{\varphi(x) \,\big|\, x \in \mathrm{co}(A)\big\} \;\geq\; \sup\big\{\varphi(x) \,\big|\, x \in c\ell\big(\mathrm{co}(A)\big)\big\}$$

The first inequality follows from linearity:

$$\varphi\left(\sum_{k=1}^{n} \lambda_k x_k\right) \;=\; \sum_{k=1}^{n} \lambda_k \varphi(x_k) \;\leq\; \sup_{k=1,\ldots,n} \varphi(x_k) \;\leq\; \sup_{x\in A} \varphi(x)$$

for any convex combination of $x_1, \ldots, x_k \,\in\, A$.

The second inequality follows from continuity according to (Denkowski et al., 2003, Theorem 1.1.29):

$$\varphi(x_0) \;=\; \lim_{\gamma\in D} \varphi(x_\gamma) \;\leq\; \sup_{x\in\mathrm{co}(A)} \varphi(x)$$

for every accumulation point x_0 of $\mathrm{co}(A)$. □

Lemma 8.30 *Let V_1 and V_2 be L-spaces. Let $B_1 \subset V_1$ be a band in V_1 and $B_2 \subset V_2$ be a band in V_2 such that $B_2 \neq \{0\}$. Then:*

a) *Every transition $\tilde{\sigma} : B_1 \to V_2$ can be extended to a transition $\sigma : V_1 \to V_2$ such that $\sigma(b_1) = \tilde{\sigma}(b_1)$ for every $b_1 \in B_1$.*

b) *For every transition $\sigma : V_1 \to V_2$, there is a transition $\tilde{\sigma} : V_1 \to B_2$ such that*

$$\sigma(x_1) \,\in\, B_2\,, \quad x_1 \in V_1 \qquad \Rightarrow \qquad \tilde{\sigma}(x_1) = \sigma(x_1) \tag{8.8}$$

Proof:

a) Fix any $\tilde{x}_2 \in V_2$ such that $x_2 \geq 0$ and $\|x_2\| = 1$. Then,

$$\rho(x_1) \;=\; \big(\|x_1^+\| - \|x_1^-\|\big) \cdot \tilde{x}_2 \qquad \forall\, x_1 \in V_1$$

defines a transition $\rho : V_1 \longrightarrow V_2$. Let π_1 be the projection of V_1 onto the band B_1 and $\pi_1^\perp$ be the projection of V_1 onto the band $B_1^\perp$. Some simple calculations show that

$$\sigma(x_1) \;=\; \tilde{\sigma} \circ \pi_1(x_1) \,+\, \rho \circ \pi_1^\perp(x_1)$$

defines an extension of $\tilde{\sigma}$ to a transition $\sigma : V_1 \longrightarrow V_2$.

b) Let π_2 be the projection of V_2 onto the band B_2 and $\pi_2^\perp$ be the projection of V_2 onto the band $B_2^\perp$. Fix any $\tilde{b}_2 \in B_2$ such that $\tilde{b}_2 \geq 0$ and $\|\tilde{b}_2\| = 1$. Then,

$$\tilde{\pi}_2(x_2) \;=\; \pi_2(x_2) + \big(\|\pi_2^\perp(x_2^+)\| - \|\pi_2^\perp(x_2^-)\|\big) \cdot \tilde{b}_2$$

defines a transition $\tilde{\pi}_2 : V_2 \longrightarrow B_2$ such that $\tilde{\pi}_2(b_2) \;=\; b_2$ for every $b_2 \in B_2$. Finally, $\tilde{\sigma} \;=\; \tilde{\pi}_2 \circ \sigma$ defines a transition $\tilde{\sigma}_2 : V_1 \longrightarrow B_2$ which fulfills (8.8).

□

Lemma 8.31 *Let Ω_1 and Ω_2 be sets with algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. Let*

$$\sigma : \;\; \mathrm{ba}(\Omega_1, \mathcal{A}_1) \;\longrightarrow\; \mathrm{ba}(\Omega_2, \mathcal{A}_2)$$

be a restricted randomization. Then, there is a finitely additive Markov kernel

$$\tau : \;\; \Omega_1 \times \mathcal{A}_2 \;\longrightarrow\; \mathbb{R}\,, \qquad (\omega_1, A_2) \;\mapsto\; \tau_{\omega_1}(A_2)$$

such that

$$\tau(\omega_1, A_2) \;=\; \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2}(\omega_1) \cdot \delta_{\tilde{\omega}_2}(A_2) \qquad \forall\, \omega_1 \in \Omega_1\,, \quad A_2 \in \mathcal{A}_2$$

where $\tilde{\Omega}_2 \,\subset\, \Omega_2$ is a finite set,

$$\alpha_{\tilde{\omega}_2} \geq 0\,, \;\; \alpha_{\tilde{\omega}_2} \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \;\; \forall\, \tilde{\omega}_2 \in \tilde{\Omega}_2 \qquad \textit{and} \qquad \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2} \equiv 1 \tag{8.9}$$

and, in addition,

$$\forall\, \tilde{\omega}_2 \in \tilde{\Omega}_2 \;\; \exists\, \tilde{A}_2 \in \mathcal{A}_2 \quad \textit{such that} \quad \tilde{A}_2 \cap \tilde{\Omega}_2 \;=\; \{\tilde{\omega}_2\} \tag{8.10}$$

Proof: It only has to be shown that $\tilde{\Omega}_2$ can be chosen in the definition of restricted randomizations in such a way that (8.10) is additionally fulfilled.

According to the definition of restricted randomizations, there is a a finitely additive Markov kernel

$$\tau : \;\; \Omega_1 \times \mathcal{A}_2 \;\longrightarrow\; \mathbb{R}\,, \qquad (\omega_1, A_2) \;\mapsto\; \tau_{\omega_1}(A_2)$$

such that

$$\tau(\omega_1, A_2) \;=\; \sum_{\tilde{\omega}_2' \in \tilde{\Omega}_2'} \beta_{\tilde{\omega}_2'}(\omega_1) \cdot \delta_{\tilde{\omega}_2'}(A_2) \qquad \forall\, \omega_1 \in \Omega_1\,, \quad A_2 \in \mathcal{A}_2$$

where $\tilde{\Omega}_2' \subset \Omega_2$ is a finite set,

$$\beta_{\tilde{\omega}_2'} \geq 0\,, \quad \beta_{\tilde{\omega}_2'} \in \mathcal{L}_\infty(\Omega_2, \mathcal{A}_2) \quad \forall\, \tilde{\omega}_2' \in \tilde{\Omega}_2' \qquad \text{and} \qquad \sum_{\tilde{\omega}_2' \in \tilde{\Omega}_2'} \beta_{\tilde{\omega}_2'} \equiv 1$$

According to finiteness of $\tilde{\Omega}_2'$, it is easy to see that there is a subset $\tilde{\Omega}_2 \subset \tilde{\Omega}_2'$ and a family of sets

$$\tilde{A}_{\tilde{\omega}_2} \in \mathcal{A}_2\,, \qquad \tilde{\omega}_2 \in \tilde{\Omega}_2$$

such that

- the sets $\tilde{A}_{\tilde{\omega}_2}$ are disjoint for $\tilde{\omega}_2 \in \tilde{\Omega}_2$,
- $\tilde{\omega}_2 \in \tilde{A}_{\tilde{\omega}_2}$ for every $\tilde{\omega}_2 \in \tilde{\Omega}_2$
- for every $\tilde{\omega}_2' \in \tilde{\Omega}_2'$ there is some $\tilde{\omega}_2 \in \tilde{\Omega}_2$ such that $\tilde{\omega}_2' \in \tilde{A}_{\tilde{\omega}_2}$
- $\tilde{\omega}_2' \in \tilde{A}_{\tilde{\omega}_2}$ implies that the following assertion is valid for every $A_2 \in \mathcal{A}_2$:

$$\tilde{\omega}_2 \in A_2 \qquad \Leftrightarrow \qquad \tilde{\omega}_2' \in A_2 \tag{8.11}$$

To say it in other words: *The elements of* $\tilde{\Omega}_2'$ *are separated by some sets* $\tilde{A}_{\tilde{\omega}_2} \in \mathcal{A}_2$ *as far as possible. If two elements of* $\tilde{\Omega}_2'$ *cannot be separated by* $\mathcal{A}_2$*, one element is too much and this redundant element is thrown away.*

Next, put

$$\alpha_{\tilde{\omega}_2} := \sum_{\tilde{\omega}_2' \in \tilde{A}_{\tilde{\omega}_2} \cap \tilde{\Omega}_2'} \beta_{\tilde{\omega}_2'} \tag{8.12}$$

Then, (8.9) and (8.10) are fulfilled. Furthermore, (8.11) and (8.12) imply

$$\tau(\omega_1, A_2) \;=\; \sum_{\tilde{\omega}_2' \in \tilde{\Omega}_2'} \beta_{\tilde{\omega}_2'}(\omega_1) \cdot \delta_{\tilde{\omega}_2'}(A_2) \;=\; \sum_{\tilde{\omega}_2 \in \tilde{\Omega}_2} \alpha_{\tilde{\omega}_2}(\omega_1) \cdot \delta_{\tilde{\omega}_2}(A_2)$$

for every $\omega_1 \in \Omega_1$ and $A_2 \in \mathcal{A}_2$. □

Lemma 8.32 *Assume that* S *is a probability charge on* $(\mathcal{U}, \mathcal{C})$ *so that* $S[\iota_\theta] = \frac{1}{n}\ \ \forall\, \theta \in \Theta$ *where* $\iota : \mathcal{U} \to \mathbb{R}$ *denotes the projection of* u *onto the* θ*-component* u_θ *of* u *. Then,* $S_\theta :\ h \mapsto S[n\iota_\theta h]$ *defines a precise model* $(S_\theta)_{\theta\in\Theta}$ *on* $(\mathcal{U}, \mathcal{C})$ *and*

$$\inf_{\rho \in \mathcal{T}_*(\mathcal{U}, \mathbb{D})} R\big((S_\theta)_\theta, \rho, W\big) = S\big[K(W)\big] \tag{8.13}$$

for every decision space $(\mathbb{D}, \mathcal{D})$ *and every loss function*

$$W \ :\ \Theta \times \mathbb{D} \ \to\ \mathbb{R}\,, \qquad (\theta, t) \ \mapsto\ W_\theta(t)\,; \qquad (W_\theta)_{\theta\in\Theta} \in \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$$

$K\big(W\big)$ *is defined as in* (4.7)*.*

Proof: Obviously, $(S_\theta)_{\theta\in\Theta}$ is a precise model on $(\mathcal{U}, \mathcal{C})$. Statement (8.13) is proven by two steps:

[1] Let $\hat{W} : (\theta, t) \mapsto \hat{W}_\theta(t)$ be a loss function such that each $\hat{W}_\theta \in \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ is a simple function. Since Θ is finite, there is a finite subset $\hat{D} := \{t_1, \ldots, t_m\} \subset \mathbb{D}$ so that

$$\Big\{ \big(\hat{W}_\theta(t)\big)_{\theta\in\Theta} \ \Big| \ t \in \hat{D} \Big\} \ = \ \Big\{ \big(\hat{W}_\theta(t)\big)_{\theta\in\Theta} \ \Big| \ t \in \mathbb{D} \Big\}$$

Let the elements of the set A be the families $(\alpha_t)_{t\in D} \subset \mathcal{L}_\infty(\mathcal{U}, \mathcal{C})$ where D is a finite subset of $\mathbb{D}$, $\alpha_t \geq 0 \ \ \forall\, t \in D$ and $\sum_{t\in D} \alpha_t \equiv 1$.
Put $\Gamma_t(u) = \sum\limits_{\theta\in\Theta} n\pi_\theta \hat{W}_\theta(t) \iota_\theta(u)$, thus $\inf\limits_{\tau\in\mathbb{D}} \Gamma_\tau = K(\hat{W})$.
For $j \in \{1, \ldots, m\}$, let V_j be the set of elements $u \in \mathcal{U}$ so that $\Gamma_{t_j}(u) = \inf_{\tau\in\mathbb{D}} \Gamma_\tau(u)$,

$$U_j \ := \ V_j \ \setminus \ \Big(\bigcup_{l=1}^{j-1} V_l \Big) \qquad \text{and} \qquad \hat{\alpha}_{t_j} \ = \ I_{U_j}, \qquad j = 1, \ldots, m$$

Note that $U_j \in \mathcal{C}$. The definition of $\{t_1, \ldots, t_m\}$ ensures that $(U_j)_{j=1,\ldots,m}$ is a partition of $\mathcal{U}$. Hence, $\sum_{t\in\hat{D}} \hat{\alpha}_t \equiv 1$ and $(\hat{\alpha}_t)_{t\in\hat{D}} \in A$. Furthermore,

$$\sum_{t\in\hat{D}} \hat{\alpha}_t(u)\Gamma_t(u) \ = \ \inf_{\tau\in\mathbb{D}} \Gamma_\tau(u) \tag{8.14}$$

Let $\hat{\rho}$ be the restricted randomization which corresponds to $(\hat{\alpha}_t)_{t\in\hat{S}} \in A$. Then,

$$\sum_{\theta\in\Theta} \pi_\theta \hat{\rho}(S_\theta)[\hat{W}_\theta] \ \overset{(8.14)}{=} \ \int \inf_{\tau\in\mathbb{D}} \Gamma_\tau(u)\, S(du) \ = \ S\big[K\big(\hat{W}\big)\big] \tag{8.15}$$

So, (8.13) follows from (8.15) and

$$\begin{aligned}
\inf_{\rho\in\mathcal{T}_*(\mathcal{U},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \rho(S_\theta)[\hat{W}_\theta] \ &\overset{\text{Prop. 4.1}}{=} \ \inf_{(\alpha_t)_{t\in D}\in A} \sum_{\theta\in\Theta} \pi_\theta S_\theta \Big[\sum_{t\in D} \hat{W}_\theta(t)\alpha_t \Big] = \\
&= \ \inf_{(\alpha_t)_{t\in D}\in A} \int \sum_{t\in D} \alpha_t(u)\Gamma_t(u)\, S(du) \ = \\
&\geq \ \inf_{(\alpha_t)_{t\in S}\in A} \int \inf_{\tau\in\mathbb{D}} \Gamma_\tau(u) \underbrace{\sum_{t\in D} \alpha_t(u)}_{=1} \ S(du) \ = \ \int \inf_{\tau\in\mathbb{D}} \Gamma_\tau(u)\, S(du)
\end{aligned}$$

[2] Fix any $\varepsilon > 0$. Then, for every $\theta \in \Theta$, there is a simple function $\hat{W}_\theta \in \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$ so that $\hat{W}_\theta - \varepsilon \ \leq \ W_\theta \ \leq \ \hat{W}_\theta + \varepsilon \ \ \ \forall\, \theta \in \Theta$; cf. (2.6). That is, $\hat{W} : (\theta, t) \mapsto \hat{W}_\theta(t)$ is a loss function as in [1]. Hence,

$$\begin{aligned}
\inf_{\rho\in\mathcal{T}_*(\mathcal{U},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \rho(S_\theta)[W_\theta] \ &\leq \ \Big(\inf_{\rho\in\mathcal{T}_*(\mathcal{U},\mathbb{D})} \sum_{\theta\in\Theta} \pi_\theta \rho(S_\theta)[\hat{W}_\theta] \Big) + \varepsilon \ = \\
&\overset{[1]}{=} \ S\Big[K(\hat{W})\Big] + \varepsilon \ = \ S\Big[\inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta \hat{W}_\theta(\tau)\iota_\theta \Big] + \varepsilon \ \leq \\
&\leq \ S\Big[\inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\iota_\theta \Big] + 2\varepsilon \ = \ S\Big[K(W)\Big] + 2\varepsilon
\end{aligned}$$

and, analogously, $\inf\limits_{\rho\in\mathcal{T}(\mathcal{U},\mathbb{D})} \sum\limits_{\theta\in\Theta} \pi_\theta \rho(S_\theta)[W_\theta] \ \geq \ S\Big[K(W)\Big] - 2\varepsilon$.
Since $\varepsilon > 0$ was arbitrarily chosen, (8.13) follows. □

Bibliography

P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit.* Vandenhoeck & Ruprecht, Göttingen, 1998.

T. Augustin. Neyman-Pearson testing under interval probability by globally least favorable pairs reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105:149–173, 2002.

T. Augustin. On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view: a cautionary note on updating imprecise priors. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 31–45. Carleton Scientific, Waterloo, 2003.

T. Augustin. Optimal decisions under complex uncertainty – basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. *ZAMM. Zeitschrift für Angewandte Mathematik und Mechanik. Journal of Applied Mathematics and Mechanics*, 84(10-11):678–687, 2004.

H. Bauer. *Probability theory.* Walter de Gruyter & Co., Berlin, 1996.

H. Bauer. *Measure and integration theory.* Walter de Gruyter & Co., Berlin, 2001.

V. Baumann. Eine parameterfreie Theorie der ungünstigsten Verteilungen für das Testen von Hypothesen. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 11: 41–60, 1968.

T. Bednarski and M. Nowak. Robustness and efficiency of sasieni-type estimators in the cox model. *J. Stat. Plann. Inference*, 115:261–272, 2003.

E.T. Bell. *The development of mathematics.* Dover Publications Inc., New York, 1992. Reprint of the second edition.

J.O. Berger. *Statistical decision theory and Bayesian analysis.* Springer-Verlag, New York, second edition, 1985.

J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-3):123–150, 2005.

K. P. S. Bhaskara Rao and M. Bhaskara Rao. *Theory of charges.* Academic Press Inc., New York, 1983. A study of finitely additive measures.

M. Bickis and U. Bickis. Predicting the next pandemic: An exercise in imprecise hazards. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 41–46. SIPTA, Prague, 2007.

P. Billingsley. *Convergence of probability measures.* John Wiley & Sons Inc., New York, 1968.

D. Blackwell. Comparison of experiments. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 1951. University of California Press.

D. Blackwell. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24:265–272, 1953.

D.V. Budescu, R. Lempert, S. Broomell, and K. Keller. Aided and unaided decision making with imprecise probabilities. *Risk Analysis*, 2008. in revision.

A. Buja. Simultaneously least favorable experiments. I. Upper standard functionals and sufficiency. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(3): 367–384, 1984.

Vašek Chvátal. *Linear programming.* W. H. Freeman and Company, New York, 1983.

C. Constantinescu, W. Filter, and K. Weber. *Advanced integration theory.* Kluwer Academic Publishers, Dordrecht, 1998. With the collaboration of Alexia Sontag.

I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In G. De Cooman, F.G. Cozman, S. Moral, and P. Walley, editors, *ISIPTA '99, Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications, held at the Universiteit Gent, Ghent, Belgium, 29 June - 2 July 1999*, pages 121–130, 1999.

F. Cozman and L. Chrisman. Learning convex sets of probability from data. Technical Report CMU-RI-TR-97-25, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 1997.

G. de Cooman. Integration and conditioning in numerical possibility theory. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):87–123, 2001. Representations of uncertainty.

G. de Cooman and P. Walley. A possibilistic hierarchical model for behaviour under uncertainty. *Theory and Decision*, 52(4):327–374, 2002.

G. de Cooman, J. Vejnarová, and M. Zaffalon, editors. *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, Prague, 2007. SIPTA.

G. de Cooman, E. Miranda, and E. Quaeghebeur. Exchangeable lower previsions. submitted, 2008.

F. Delbaen. Coherent risk measures on general probability spaces. In *Advances in finance and stochastics*, pages 1–37. Springer, Berlin, 2002.

Z. Denkowski, S. Migórski, and N.S. Papageorgiou. *An introduction to nonlinear analysis: theory.* Kluwer Academic Publishers, Boston, 2003.

D.L. Donoho and R.C. Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

R.M. Dudley. *Real analysis and probability.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

N. Dunford and J.T. Schwartz. *Linear Operators. I. General Theory.* Wiley-Interscience Publishers, New York, 1958.

D. Ellsberg. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39:42–47, 1953.

P.I. Fierens and T.L. Fine. Towards a chaotic probability model for frequentist probability: The univariate case. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 245–259. Carleton Scientific, Waterloo, 2003.

H. Föllmer and A. Schied. Robust preferences and convex measures of risk. In *Advances in finance and stochastics*, pages 39–56. Springer, Berlin, 2002.

H. Föllmer and A. Schied. *Stochastic finance.* Walter de Gruyter & Co., Berlin, extended edition, 2004. An introduction in discrete time.

H. Föllmer, A. Schied, and S. Weber. Robust preferences and robust portfolio choice, 2007. submitted.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

R. Hable. Data-based decisions under imprecise probability and least favorable models. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 203–212. SIPTA, Prague, 2007.

R. Hable. *imprProbEst: Minimum distance estimation in an imprecise probability model*, 2008a. Contributed R-Package on CRAN, Version 1.0, 2008-10-23; maintainer Hable, R.

R. Hable. Data-based decisions under imprecise probability and least favorable models. *International Journal of Approximate Reasoning*, 2008b. in press.

R. Hable, P. Ruckdeschel, and H. Rieder. Optimal robust influence functions in semiparametric regression. *Journal of Statistical Planning and Inference*, 2008. in revision.

R. Hafner. Konstruktion robuster Teststrategien. In S. Schach and G. Trenkler, editors, *Data analysis and statistical inference*, pages 145–160. Eul, Bergisch Gladbach, 1992.

J. Hall, G. Fu, and J. Lawry. Imprecise probabilities of climate change: aggregation of fuzzy scenarios and model uncertainties. *Climatic Change*, 81:265–281, 2007.

O. Hamouda and J.C.R. Rowley. *Paradoxes, Ambiguity and Rationality.* Edward Elgar, Cheltenham, 1997.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics.* John Wiley & Sons Inc., New York, 1986.

H. Held, T. Augustin, and E. Kriegler. Bayesian learning for a class of priors with prescribed marginals. *International Journal of Approximate Reasoning*, 49(1):212–233, 2008.

H. Heyer. Erschöpftheit und Invarianz beim Vergleich von Experimenten. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 12:21–55, 1969.

H. Heyer. *Mathematische Theorie statistischer Experimente.* Springer-Verlag, Berlin, 1973.

J. Hoffmann-Jørgensen. *Probability with a view toward statistics. Vol. I.* Chapman & Hall, New York, 1994a.

J. Hoffmann-Jørgensen. *Probability with a view toward statistics. Vol. II.* Chapman & Hall, New York, 1994b.

P.J. Huber. A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 36:1753–1758, 1965.

P.J. Huber. *Robust statistics.* John Wiley & Sons Inc., New York, 1981.

P.J. Huber. *Robust statistical procedures. 2nd ed.* CBMS-NSF Regional Conference Series in Applied Mathematics. 68. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics., 1997.

P.J. Huber and V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, 1:251–263, 1973.

M. Hutter. Practical robust estimators for the imprecise dirichlet model. *International Journal of Approximate Reasoning*, 2008. in press.

B. Jansen, J.J. de Jong, C. Roos, and T. Terlaky. Sensitivity analysis in linear programming: Just be careful! *European Journal of Operational Research*, 101(1):15–28, 1997.

J. Jurečková and J. Picek. *Robust statistical methods with R.* Chapman & Hall/CRC, Boca Raton, FL, 2006.

J. Jurečková and P.K. Sen. *Robust statistical procedures.* John Wiley & Sons Inc., New York, 1996. Asymptotics and interrelations, A Wiley-Interscience Publication.

D. Kikuti, F.G. Cozman, and C.P. de Campos. Partially ordered preferences in decision trees: computing strategies with imprecision in probabilities. In R. Brafman and U. Junker, editors, *Multidisciplinary IJCAI-05 Workshop on Advances in Preference Handling, Edinburgh, United Kingdom*, 2005.

M. Kohl. *Numerical Contributions to the Asymptotic Theory of Robustness.* PhD thesis, Universität Bayreuth, 2005.

A.N. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Julius Springer, Berlin, 1933.

A. N. Kolmogorov. *Foundations of the theory of probability.* Chelsea Publishing Co., New York, 1956. Translation edited by Nathan Morrison, with an added bibliography by A. T. Bharucha-Reid.

H. König. *Measure and integration.* Springer-Verlag, Berlin, 1997.

Volker Krätschmer. Robust representation of convex risk measures by probability measures. *Finance Stoch.*, 9(4):597–608, 2005.

E. Kriegler. *Imprecise Probability Analysis for Integrated Assessment of Climate Change.* PhD thesis, Universität Potsdam, 2005.

E. Kriegler and H. Held. Climate projections for the 21st century using random sets. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 345–360. Carleton Scientific, Waterloo, 2003.

J.B. Lasserre. Weak convergences of probability measures: a uniform principle. *Proceedings of the American Mathematical Society*, 126(10):3089–3096, 1998.

L. Le Cam. Sufficiency and approximate sufficiency. *Annals of Mathematical Statistics*, 35:1419–1455, 1964.

L. Le Cam. *Asymptotic methods in statistical decision theory.* Springer-Verlag, New York, 1986.

F. Liese and K. Miescke. *Statistical decision theory.* Springer, New York, 2008.

D.G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons Inc., New York, 1969.

F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

A. Marazzi. *Algorithms, routines, and S functions for robust statistics.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1993. The FORTRAN library ROBETH with an interface to S-PLUS, With the collaboration of Johann Joss and Alex Randriamiharisoa, With a separately available computer disk.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust statistics.* John Wiley & Sons Ltd., Chichester, 2006. Theory and methods.

D. McAllester and P. Myllymäki, editors. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, Helsinki, Finland, 2008.

P. McCullagh and J.A. Nelder. *Generalized linear models.* Chapman & Hall, London, 1983.

P.W. Millar. Robust estimation via minimum distance methods. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):73–89, 1981.

E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 46(1):188–225, 2007.

C.H. Müller. *Robust planning and analysis of experiments*. Springer-Verlag, New York, 1997.

M. Obermeier and T. Augustin. Lucenos discretization methods and its application in decision making under ambiguity. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 327–336. SIPTA, Prague, 2007.

F. Österreicher. On the construction of least favourable pairs of distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 43:49–55, 1978.

Ö. Öztürk and T.P. Hettmansperger. Simultaneous robust estimation of location and scale parameters: a minimum-distance approach. *The Canadian Journal of Statistics.*, 26(2):217–229, 1998.

W.C. Parr and W.R. Schucany. Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624, 1980.

R. Pelessoni and P. Vicig. Uncertainty modelling and conditioning with convex imprecise previsions. *International Journal of Approximate Reasoning*, 39(2–3):297–319, 2005.

E. Quaeghebeur and G. de Cooman. Imprecise probability models for inference in exponential families. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburg*, pages 287–296. SIPTA, Manno, 2005.

L.C. Rêgo and T.L. Fine. Estimation of chaotic probabilities. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburg*, pages 297–305. SIPTA, Manno, 2005.

H. Rieder. Robuste Tests. Diplomarbeit, Albert-Ludwigs-Universität Freiburg, 1974.

H. Rieder. Least favorable pairs for special capacities. *The Annals of Statistics*, 5:909–921, 1977.

H. Rieder. *Robust asymptotic statistics*. Springer-Verlag, New York, 1994.

P. Sasieni. Some new estimators for Cox regression. *Ann. Statist.*, 21:1721–1759, 1993.

H.H. Schaefer. *Banach lattices and positive operators*. Springer-Verlag, Berlin-Heidelberg-New York, 1974.

M.J. Schervish, T. Seidenfeld, J.B. Kadane, and I. Levi. Extensions of expected utility theory and some limitations of pairwise comparisons. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 496–510. Carleton Scientific, Waterloo, 2003.

A. Schied. Risk measures and robust optimization problems. *Stochastic Models*, 22(4): 753–831, 2006.

J. Shao. *Mathematical statistics.* Springer-Verlag, New York, second edition, 2003.

W.F. Sheppard. On the calculation of the most probable values of frequency-constants for data arranged according to equidistant divisions of a scale. *Proceedings of the London Mathematical Society*, 29:353–380, 1898.

D. Škulj. Jeffrey's conditioning rule in neighbourhood models. *International Journal of Approximate Reasoning*, 42(3):192–211, 2006.

H. Strasser. *Mathematical theory of statistics.* Walter de Gruyter & Co., Berlin, 1985.

H. Strasser. Review of the book: *Asymptotic methods in statistical decision theory.* Zentralblatt MATH, Springer-Verlag, 2008.
`http://www.zentralblatt-math.org/zmath/en/advanced/` .

E. Torgersen. *Comparison of statistical experiments.* Cambridge University Press, Cambridge, 1991.

M.C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45:17–29, 2007.

M.C.M. Troffaes. Finite approximations to coherent choice. *International Journal of Approximate Reasoning*, 2008. in press.

J.W. Tukey. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pages 448–485. Stanford Univ. Press, Stanford, Calif., 1960.

L.V. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburg*, pages 349–358. SIPTA, Manno, 2005.

A. van der Vaart. The statistical work of Lucien Le Cam. *The Annals of Statistics*, 30 (3):631–682, 2002. Dedicated to the memory of Lucien Le Cam.

A. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes. With applications to statistics.* Springer, New York, 1996.

A.W. van der Vaart. *Asymptotic statistics.* Cambridge University Press, Cambridge, 1998.

V.S. Varadarajan. Measures on topological spaces. *American Mathematical Society Translations*, 48:161–228, 1965.

B. Vidakovic. Γ-minimax: a paradigm for conservative robust Bayesians. In *Robust Bayesian analysis*, pages 241–259. Springer-Verlag, New York, 2000.

P. Vigic. Financial risk measurement with imprecise probabilities. *International Journal of Approximate Reasoning*, 49(1):159–174, 2008.

A. Wald. *Statistical Decision Functions.* John Wiley & Sons Inc., New York, 1950.

P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman & Hall, London, 1991.

P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):3–57, 1996. With discussion and a reply by the author.

G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized bayesian inference. *Journal of Statistical Theory and Practice: Special Issue on Imprecision*, 2008. in press.

K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000. Reasoning with imprecise probabilities (Ghent, 1999).

K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica, Heidelberg, 2001.

K. Weichselberger and T. Augustin. Analysing Ellsberg's paradox by means of interval-probabilty. In R. Galata and H. Küchenhoff, editors, *Econometrics in theory and practice. Festschrift for Hans Schneeweiß*, pages 291–304. Physica, 1998.

K. Weichselberger and T. Augustin. On the symbiosis of two concepts of conditional interval probability. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 608–630. Carleton Scientific, Waterloo, 2003.

R.R. Wilcox. *Introduction to robust estimation and hypothesis testing.* Academic Press Inc., San Diego, CA, 1997.

H. Witting. *Mathematische Statistik. I.* B.G. Teubner, Stuttgart, 1985.

R.R. Yager, J. Kacprzyk, and M. Fedrizzi, editors. *Advances in the Dempster-Shafer theory of evidence.* Wiley, 1994.

G.L. Yang. Lucien Le Cam 1924–2000. *The Annals of Statistics*, 30(3):617–630, 2002. Dedicated to the memory of Lucien Le Cam.

L.A. Zadeh and J. Kacprzik, editors. *Fuzzy logic for the management of uncertainty.* Wiley, New York, 1992.

List of Symbols

Index

Lebenslauf

Robert Hable

28.08.1981	geboren in Landshut
Familienstand:	verheiratet
seit 04/2008	**Universität Bayreuth** Wissenschaftlicher Mitarbeiter am Mathematischen Institut
10/2006 - 09/2008	**LMU München** Promotionsstudent am Institut für Statistik Betreuung: Prof. Dr. T. Augustin
10/2007 - 11/2007	Wissenschaftlicher Mitarbeiter am Institut für Statistik
05/2006 - 09/2006	**LMU München** Wissenschaftliche Hilfskraft am Institut für Statistik
10/2001 - 03/2006	**Univesität Bayreuth** Studium *Diplom Mathematik*, mit Nebenfach VWL studentische Hilfskraft am Mathematischen Institut und an der Rechts- und Wirtschaftswissenschaftlichen Fakultät Abschluss: Diplom Mathematiker (Uni.) Note: 1,0
09/1992 - 06/2001	**Maximilian-von-Montgelas-Gymnasium Vilsbiburg** Abschluss: Abitur Note: 1,4

Stipendiat des Cusanuswerks

Förderung im Studium:	04/2004 - 03/2006
Promotionsförderung:	10/2006 - 04/2008

Gewinner des *IJAR Student Award* des *International Journal of Approximate Reasoning*, 2007