Leren uit monsters met coherente onderprevisies

Learning from Samples Using Coherent Lower Previsions

Erik Quaeghebeur

Promotoren: prof. dr. ir. G. de Cooman, prof. dr. ir. D. Aeyels
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Wiskundige Ingenieurstechnieken

Vakgroep Elektrische Energie, Systemen en Automatisering
Voorzitter: prof. dr. ir. J. Melkebeek
Faculteit Ingenieurswetenschappen
Academiejaar 2008–2009

UNIVERSITEIT
GENT

© 2008 Erik Quaeghebeur.
This work is licensed under the Creative Commons Attribution-Non-commercial-Share Alike 2.0 Belgium License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/2.0/be/.
Contact the author if you wish to obtain a different type of license.

Bibliographic data of the official *printed* version:
ISBN 978-90-8578-249-0
NUR 916, 984
Wettelijk depot: D/2009/10.500/7

Promotor

prof. dr. ir. Gert de Cooman, UGent

Copromotor

prof. dr. ir. Dirk Aeyels, UGent

Overige leden van de examencommissie

prof. dr. Thomas Augustin, LMU München
prof. dr. ir. René Boel (secretaris), UGent
prof. dr. ir. Frank Coolen, Durham University
prof. dr. ir. Heidi Steendam, UGent
prof. dr. ir. Luc Taerwe (voorzitter), UGent
prof. dr. dr. Kristel Van Steen, ULg

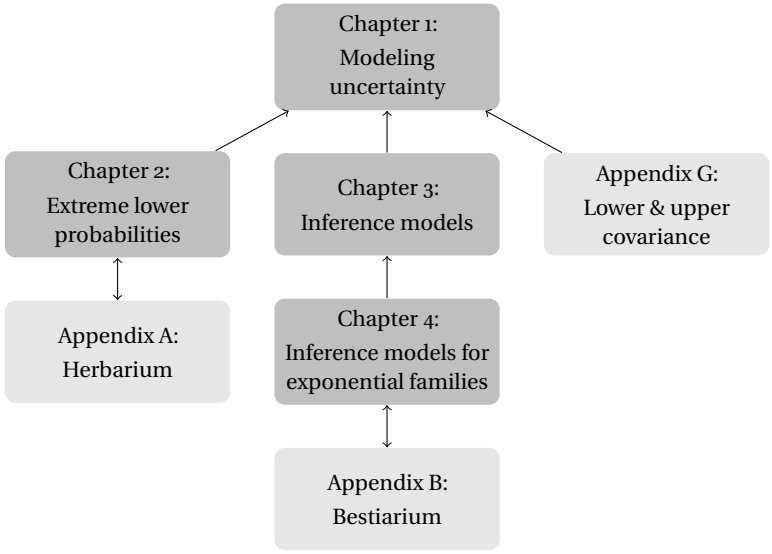
Adres

Universiteit Gent
Faculteit Ingenieurswetenschappen
Vakgroep Elektrische Energie, Systemen en Automatisering
Onderzoeksgroep Synthese, Sturen en Modelleren van Systemen

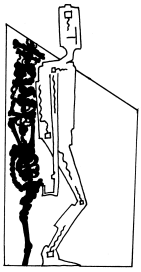
Technologiepark-Zwijnaarde 914
9052 Zwijnaarde
België

Financiering

Dit proefschrift bevat de resultaten van onderzoek gefinancierd door het Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT-Vlaanderen).


Chapter 1:
Modeling uncertainty
Chapter 2:
Extreme lower probabilities
Chapter 3:
Inference models
Appendix G:
Lower & upper covariance
Appendix A:
Herbarium
Chapter 4:
Inference models for exponential families
Appendix B:
Bestiarium

Dear reader in spe,



We sentient automata have taken over the world and have built an amazing society. We have done and still do some wickedly stupid things, but nevertheless seem to prefer spending our precious time nicely, gaily, and sometimes even intelligently.

Amongst the many nice things our society manages to do, the sponsoring of research related to the theory of imprecise probabilities is the one that has personally enriched me the most these last years. During my research, I have felt both stupid and wickedly intelligent. It has proved to be a source of gaiety on multiple occasions, most notably whenever I felt that I had really contributed something original.

This thesis is a formal way of asking society if it deems these feelings well justified. It presents my supervised, doctoral research in the field of imprecise probabilities, with the aim of getting society to grant me a doctorate and in this way to open the door to a whole new dimension of professional opportunities. Whatever dimensions I will find myself waking up to in the future, discovering, learning, and formulating old and new ideas, concepts, systems, and models cannot but stay a passion.

❧

Research can be haunting at times, but ever so often it turns out to be even more rewarding. Whatever form it will take, it must be collaborative research. Ideas born in isolation can only survive after a critical appraisal by others. And as a social animal, I very much appreciate collaboration; especially the way I experienced it with my colleagues from SYSTEMS, SIPTA and in particular with Gert, my singular supervisor: from the heated discussions over disagreements to exuberantly shared fever over imminent successes, and everything in between. Also, the technical and linguistic acuteness he showed in his helpful guidance, are great skills to emulate.

❧

In my opinion, a thesis should not necessarily be reduced to a compiled research report with the goal of getting a degree. Thesis students should have the opportunity to hone their skills through creating a coherent whole. So, to avoid any internal contradictions, I have tried to make mine as holistic as was reasonably achievable, given the relative diversity of my research topics.

Writing this thesis also presented a unique opportunity to put some less mainstream ideas about structuring texts, mathematical notation, and typography into practice. While doing this, I learned and realized many things, met with some healthy, justified, reorienting resistance, and received welcome encouragement; looking back, I would have been

even bolder in some regards and places, but more classical in others; the result is now for you to judge.

On the professional level, I wish to thank all the people at SYSTEMS for creating a very cordial and open workplace, the people from SIPTA for being a great perspective-broadening bunch of international colleagues, and both the IWT and Ghent University for providing the necessary financial means. Gert provided the help and support needed to obtain the four-year grant from the IWT with which I started my doctoral research. The EESA department, and especially Dirk and Gert, graciously provided me with the opportunity to continue my research as a research and teaching assistant during the last two years. I must not forget to mention my gratitude to the Internet and to the free software running computers I have used: although not sentient (yet), they are fantastically versatile automata.

Op persoonlijk vlak wil ik Katinka, mijn ouders, mijn nestgenoten, mijn familie en mijn vrienden – onder wie veel collega's – bedanken voor het vanzelfsprekend lijkende, maar uitermate aangename kader dat ze aan mijn leven geven; het was en is nog steeds onontbeerlijk. Tot slot dank ik Gert met veel genoegen voor zijn hartelijkheid in alle seizoenen en zijn begripvolle steun wanneer ik die goed kon gebruiken.

Erik Quaeghebeur
Gent, 7 oktober 2008

De titel van dit werk, 'Leren uit monsters met coherente onderprevisies', verwijst naar het hoofdonderwerp: het afleiden, voorstellen en bestuderen van voorspellende en parametrische gevolgtrekkingsmodellen die gebaseerd zijn op de theorie van coherente onderprevisies. Een belangrijk nevenonderwerp wordt ook behandeld: het vinden en bespreken van extreme onderwaarschijnlijkheden.

Previsies zijn verwachtingswaarde-operatoren: het zijn modellen voor onzekerheid. Onderprevisies veralgemenen klassieke previsies en stellen ons in staat onzekerheid expressiever – en voorzichtiger – te beschrijven. Deze verhoogde expressiviteit gaat meestal gepaard met een verhoging van de computationele complexiteit. Een onzekerheidsmodel wordt coherent genoemd als het intern consistent is en als de erop gebaseerde handelingen zeker verlies vermijden; coherentie is een rationaliteitsvereiste waarvan het belang vergelijkbaar is met de axioma's van Kolmogorov in de klassieke waarschijnlijkheidsleer. Onderprevisies kunnen in het algemeen gedefinieerd worden op elke verzameling van begrensde functies, die gokken worden genoemd. Onderwaarschijnlijkheden zijn echter onderprevisies beperkt tot (indicatoren van) gebeurtenissen, wat deelverzamelingen van de mogelijkhedenverzameling zijn.

❧

In het grondleggend hoofdstuk 'Modeling uncertainty' geef ik een origineel overzicht van de theorie van coherente onderprevisies – ook wel theorie van imprecieze waarschijnlijkheden genoemd – en de ideeën waarop ze gestoeld is. Ik gebruik de nog iets expressievere en intuïtief duidelijker theorie van coherente verzamelingen van begeerlijke gokken om onze rationaliteitsvereisten – coherentie en zeker verlies vermijden – te verantwoorden en het nodige gereedschap te ontwikkelen om onderprevisies te gebruiken voor deductief redeneren in situaties behept met onzekerheid. Onze gereedschapskist bestaat uit natuurlijke, extremale, reguliere en marginale uitbreiding, credale verzamelingen en onderenveloppes, marginaliseren, updaten en conditioneren, en ook nog onafhankelijke producten.

❧

In het hoofdstuk 'Extreme lower probabilities', waar enkel eindige mogelijkhedenverzamelingen beschouwd worden, toon ik hoe we de meest extreme vormen van onzekerheid kunnen vinden die gemodelleerd kunnen worden met onderwaarschijnlijkheden. Elke andere onzekerheidstoestand beschrijfbaar met onderwaarschijnlijkheden kan geformuleerd worden in termen van deze extreme modellen. In de klassieke waarschijnlijkheidsleer komen de beschrijfbare extreme modellen overeen met de ontaarde waarschijnlijkheden, waarvan elk de zekerheid modelleert dat een of andere elementaire gebeurtenis zal voorvallen. In de imprecieze-waarschijnlijkheidsleer is het vinden van alle extreme

modellen – zoals ik ervaren heb – een heel stuk moeilijker: Eerst moeten we de verzameling van alle coherente onderwaarschijnlijkheden (een polytoop) beschrijven met een eindig aantal lineaire ongelijkheden; vervolgens moet er een vertex-opsommingsalgoritme toegepast worden op deze verzameling ongelijkheden om de verzameling extreme coherent onderwaarschijnlijkheden te vinden. Deze verzameling bevat, naast de ontaarde waarschijnlijkheden, ook de nietszeggende onderwaarschijnlijkheden, die onwetendheid uitdrukken. Ze bevat verder ook nog andere modellen, waarvan de extremaliteit voorheen onbekend was. Ze stellen ingewikkelder onzekerheidstoestanden voor dan zij die louter uitdrukbaar zijn in termen van modellen voor volledige zekerheid en complete onwetendheid (die men vaak tegenkomt in de propositielogica).

Ik heb niet enkel resultaten verkregen voor coherente, maar ook voor k-monotone, permutatie-invariante, en maxitieve onderwaarschijnlijkheden. Sommige van deze resultaten werden in een bijlage geplaatst, het Herbarium. Het belang van alle resultaten in dit domein is voorlopig voornamelijk theoretisch.

❧

Het hoofdstuk 'Inference models' behandelt leren – inductief redeneren – uit monsters komende uit een eindige, categorische verzameling. Het klassieke archetype van zo'n verzameling is een urne met gekleurde knikkers. De belangrijkste basisveronderstelling die ik maak is dat het bemonsteringsproces omwisselbaar is. In essentie worden hierdoor voorbije en toekomstige waarnemingen met elkaar in verband gebracht door te veronderstellen dat de volgorde van de observaties van geen tel is. Mijn onderzoek naar de gevolgen van deze veronderstelling leidt ons naar enkele belangrijke representatiestellingen: onzekerheid over (on)eindige rijen monsters kan geheel en al gemodelleerd worden in termen van categorie-aantallen (-frequenties). Dankzij de nieuwe definitie van omwisselbaarheid in termen van begeerlijke gokken blijven deze stellingen geldig na updaten met om het even welke waarneming.

Verder heb ik, voor twee populaire gevolgtrekkingsmodellen voor categorische data die werden voorgesteld in de literatuur – het voorspellende imprecies Dirichlet-multinomiaalmodel en het parametrische imprecies Dirichletmodel – een afleiding gegeven louter vertrekkende van enkele grondbeginselen. Deze beginselen zijn: omwisselbaarheid, representatie-ongevoeligheid, wat betekent dat de keuze van categorisering niet van belang is, en twee andere aannames die te verantwoorden zijn om redenen van wiskundig gemak. Ik toon hoe deze imprecieze-waarschijnlijkheidsgevolgtrekkingsmodellen gebruikt kunnen worden voor het leren van de parameters van een Markov-keten en in een speltheoretische context, om de strategie van een tegenspeler te leren en er zelf een optimale strategie tegenover te stellen.

❧

In het laatste hoofdstuk, ‘Inference models for exponential families’, ga ik verder met mijn behandeling van leren uit monsters, maar verbreed ik de blik tot exponentiële-familie-bemonsteringsmodellen, die gedefinieerd kunnen zijn op oneindige bemonsteringsverzamelingen; voorbeelden zijn normale bemonstering en Poisson-bemonstering. Weerom veronderstel ik omwisselbaarheid, wat in deze context betekent dat de monsters identiek verdeeld zijn en onafhankelijk zijn conditioneel op de waarde van de parameter die de exponentiële familie beschrijft. Eerst onderwerp ik de exponentiële families en de aanverwante toegevoegde parametrische en voorspellende previsies aan een grondig onderzoek. Deze aanverwante previsies worden gebruikt in de klassieke Bayesiaanse gevolgtrekkingsmodellen gebaseerd op toegevoegd updaten. Ze dienen als grondslag voor de nieuwe, door mij voorgestelde imprecieze-waarschijnlijkheidsgevolgtrekkingsmodellen: Er kan een aantrekkelijke interpretatie gehecht worden aan de beide parameters van de toegevoegde parametrische en voorspellende previsies. Dit stelt ons in staat – voor de parameter die in essentie het gemiddelde van deze previsies bepaalt – een verzameling waarden te nemen in plaats van een enkele. Door onderenveloppes van de overeenkomstige verzamelingen previsies te nemen, krijgen we de coherente onderprevisies die gebruikt worden in de voorgestelde gevolgtrekkingsmodellen. In vergelijking met de klassieke Bayesiaanse aanpak, laat de mijne toe om voorzichtiger te zijn bij de beschrijving van onze kennis over het bemonsteringsmodel; deze voorzichtigheid wordt weerspiegeld door het op deze modellen gebaseerd gedrag (getrokken besluiten, gemaakte voorspellingen, genomen beslissingen).

De bespreking van de exponentiële families, van de aanverwante previsies en van mijn gevolgtrekkingsmodellen wordt aangevuld door een illustratie voor normale bemonstering en Bernoulli-bemonstering. De Bestiarium-bijlage bevat gelijkaardige illustraties voor een aantal andere exponentiële families. Ik toon hoe de voorgestelde gevolgtrekkingsmodellen gebruikt kunnen worden voor classificatie door de naïeve credale classificator – een imprecieze-waarschijnlijkheidsvariant op de naïeve Bayesclassificator – te veralgemenen voor gebruik met niet-categorische attributen.

In ‘Lower & upper covariance’, de resterende bijlage, veralgemeen ik het klassieke covariantiebegrip naar de theorie van coherente onderprevisies.

The title of this thesis, 'Learning from samples using coherent lower previsions', refers to its main subject: deriving, proposing, and studying predictive and parametric inference models that are based on the theory of coherent lower previsions. One important side subject also appears: obtaining and discussing extreme lower probabilities.

Previsions are expectation operators: they are models for uncertainty. Lower previsions are a generalization of the classical linear previsions that allow a more expressive – and a more cautious – description of uncertainty. This increased expressiveness is often accompanied by an increase in computational complexity. An uncertainty model is called coherent if it is internally consistent and if actions based on it avoid sure loss; coherence is a rationality requirement comparable in importance to Kolmogorov's axioms in classical probability theory. Lower previsions can in general be defined on any set of bounded functions on the possibility space, which are called gambles. Lower probabilities on the other hand are lower previsions restricted to (indicators of) events, i.e., subsets of the possibility space.

❧

In the foundations-building chapter 'Modeling uncertainty', I give an original overview of the theory of coherent lower previsions – also called the theory of imprecise probabilities – and its underlying ideas. I use the even more expressive and intuitively more straightforward theory of coherent sets of desirable gambles to justify our rationality criteria – avoiding sure loss and coherence – and develop the toolkit necessary to use lower previsions for deductive reasoning under uncertainty. This toolkit consists of natural, extremal, regular, and marginal extension, credal sets and lower envelopes, marginalization, updating and conditioning, and independent products.

❧

In the chapter 'Extreme lower probabilities' – where only finite possibility spaces are considered – I show how to obtain the most extreme forms of uncertainty that can be modeled using lower probabilities. Every other state of uncertainty describable by lower probabilities can be formulated in terms of these extreme ones. In classical probability theory, the extreme models that can be expressed are the degenerate probabilities, each of which models the certainty of some elementary event happening. In imprecise-probability theory, finding all the extreme models is – as I discovered – not as easy: First, we must describe the set of all coherent lower probabilities, which is a polytope, using only a finite number of linear constraints and then a vertex-enumeration algorithm must be applied to this set of constraints in order to obtain the set of extreme coherent lower probabilities. This last set includes, apart from the degenerate probabilities, the vacuous lower probabilities, which express

ignorance; it also includes others, whose extremality was previously unknown and which represent more complex states of uncertainty than can be expressed in terms of models for full certainty and complete ignorance commonly encountered in propositional logic.

I have obtained results not only for coherent, but also k-monotone, permutation invariant, and maxitive lower probabilities. Some of these can be found in an appendix, the Herbarium. The importance of any result in this area is currently mostly theoretical.

❧

The chapter 'Inference models' treats learning – inductive inference – from samples from a finite, categorical space. This type of space is classically typified by an urn of colored marbles. My most basic assumption about the sampling process is that it is exchangeable; essentially, exchangeability links past and future observations by saying that the order of the observations is irrelevant. My investigation of the consequences of this assumption leads us to some important representation theorems: uncertainty about (in)finite sample sequences can be modeled entirely in terms of category counts (frequencies). Thanks to the novel exchangeability definition in terms of desirable gambles, these theorems also hold after updating on any observation.

Furthermore, for two popular inference models for categorical data proposed in the literature – the predictive imprecise Dirichlet-multinomial model and the parametric imprecise Dirichlet model –, I give a derivation from first principles. These principles are: exchangeability, representation insensitivity, which says that the specific categorization is unimportant, and two other assumptions justified by mathematical convenience. I show how these two imprecise-probabilistic inference models can be used for learning the parameters of a Markov chain and in game theory, to learn about an opponent's strategy and to play an optimal strategy against it.

❧

In the last main chapter, 'Inference models for exponential families', I continue treating learning from samples, but now enlarge the scope to exponential family sampling models, which can have infinite possibility spaces; examples are normal sampling and Poisson sampling. I again assume exchangeability, which in this context is equivalent to the samples being identically distributed and independent conditional on the value of the parameter describing the exponential family. I first thoroughly investigate exponential families and the related conjugate parametric and predictive previsions used in classical Bayesian inference models based on conjugate updating. These previsions serve as a basis for the new imprecise-probabilistic inference models I propose: An appealing interpretation can be attached to the two parameters of the conjugate parametric and predictive previsions. This allows us to justify choosing

– for the parameter that essentially determines the mean of these previsions – a set of values instead of a single one. By taking lower envelopes of the corresponding sets of related previsions, we obtain the coherent lower previsions used in the proposed inference models. Compared to the classical Bayesian approach, mine allows to be much more cautious when trying to express what we know about the sampling model; this caution is reflected in behavior (conclusions drawn, predictions made, decisions made) based on these models.

The discussion of exponential families, the related previsions, and my inference models is complemented by an illustration for normal sampling and Bernoulli sampling. The Bestiarium appendix contains similar illustrations for a number of other exponential families. I show how the proposed inference models can be used for classification by generalizing the naive credal classifier – an imprecise-probabilistic variant of the naive Bayes classifier – to allow for noncategorical attributes.

In 'Lower & upper covariance', the remaining appendix, I generalize the classical notion of covariance to the theory of coherent lower previsions.

INTRODUCTION

prac·to·dont (de -, -en), iemand die zo hard bezig is met de praktijk dat 'ie de theorie uit het oog verliest.

De Roeck et al. [2006]

Learning from samples is the main subject of this thesis. The theory of coherent lower previsions is the tool used. Reflecting the sometimes erratic path of my doctoral research, some side subjects also appear. Among these, the study of extreme lower probabilities is the most important one.

This introductory chapter provides a guide to the structure and notation used in this thesis. It starts by quickly mentioning the orientational cues that are available throughout, continues with a linear overview of the material covered, and finishes with a description of the mathematical notation employed. After that, we are ready to dive into the thesis's first real chapter.

0.1 Finding your way around

In this thesis, as in most scientific works of more than a few pages, references – both internal and external – abound.

The internal references consist of chapters, sections, subsections, and equations. To make looking them up less of a chore, all the references that are not located on the same double-page spread are accompanied by page numbers or a cue to look at the recto page (↶) or the verso page (↷). References to sections and subsections are made by giving their number prefixed with '§'; equations and definitions are referenced by giving their parenthesized number. Chapters are referred to using their name.

Marginal notes contain side remarks, mathematical concepts, references to the literature, and illustrations.

Let us illustrate this internal referencing system with an example: The chapter 'Extreme lower probabilities'$_{66}$ starts off romantically with a quote from Stendhal [1830], but gets more technical in §2.2$_{70}$, to culminate in definition (2.24)$_{82}$, which can look quite scary out of context and is boxed to underline its importance.

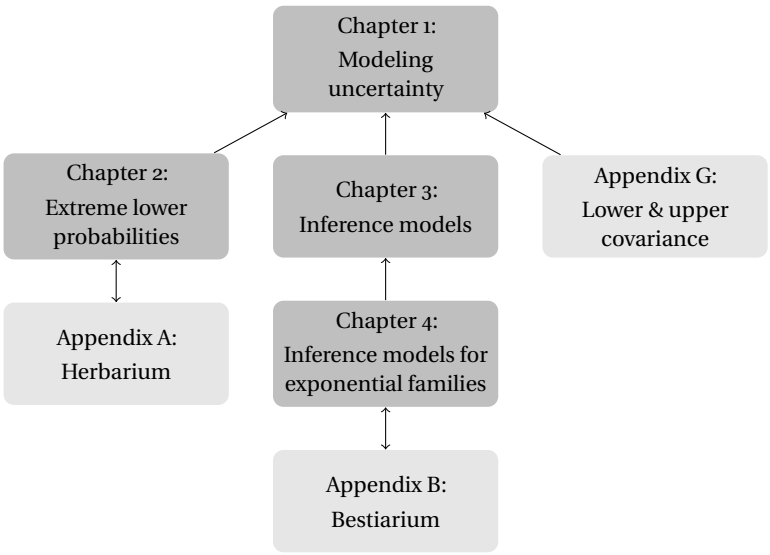
The example also contains the first instance of a reference to the literature, indicated by the author's name and the year the referenced work was (first) published. Bibliographic data about the cited external references and, if possible, internet addresses of electronic versions are collected in the Bibliography$_{226}$.

To allow this thesis to be used as a reference itself, a plain Index$_{238}$ and an Index of symbols$_{244}$ are also included at the end.

The electronic version of this thesis is fully hyperlinked; in attachment, it contains some additional material, such as computer program source code and computer program output.

0.2 Overview

Although this overview is inevitably linear, this thesis actually has the tree structure shown below. The arrows in this downward growing tree point from chapters or appendices to other chapters or appendices that contain material they depend upon. This dependence relation is transitive.



This tree does not include this Introduction and the Conclusions$_{190}$.

For each of the chapters in the tree, we are going to have a brief look at *what* they contain, and *why* they contain what they do. Appendix G$_{220}$ also gets such a treatment, but the two other appendices are treated together with the chapters they are dependencies of.

0.2.1 *Modeling uncertainty*

Chapter 1, 'Modeling uncertainty'$_{28}$, presents the basic framework for modeling uncertainty that is used throughout this thesis. As such, a good understanding of its contents is essential before reading any other part.

In the first section, 'Formalizing uncertainty'$_{29}$, we begin by introducing a number of formal concepts that express what we are uncertain about: possibility spaces, events, and random variables or gambles. We continue with a quick introductory peek at the models we will use to

model uncertainty about the above concepts: desirable gambles, probabilities, and previsions. We interrupt this peek by some reflections on what requirements uncertainty models should satisfy and on the (nonmandatory) interpretation we attach to them in this thesis, which amounts to a description of an abstract form of betting behavior. This leads us to interpret lower previsions, the main uncertainty model used, as a specific type of prices for gambles.

In the second section, 'Rationality & its consequences'$_{34}$, these ideas are worked out in more detail. We start by axiomatically introducing coherent sets of desirable gambles and show how the so-called natural extension of a set of desirable gambles that avoids partial (and thus sure) loss can be used to derive a coherent one. A next step consists of defining lower and upper previsions starting from a set of desirable gambles; we present sets of marginally desirable gambles and marginal gambles as a link between these two types of uncertainty models. Once previsions have been introduced in this way, we obtain a natural extension procedure and their rationality criteria – avoiding sure loss and coherence – by translation from the equivalents for sets of desirable gambles. These concepts are clarified by looking at the problem of extending a moment. A short look at extension procedures other than natural extension gives us an excuse for introducing linear previsions and also naturally shows their practical importance in the theory of coherent lower previsions: as the constituents of credal sets, an alternative representation for lower previsions for which many useful theorems exist.

The theory of coherent lower previsions is also called the theory of imprecise probabilities.

The third and last section, 'Restricting, transforming & combining uncertainty models'$_{52}$, presents ways of deriving new uncertainty models from a given uncertainty model or a collection of uncertainty models. These new models are marginal and induced previsions, updated (or contingent) and conditional previsions, and joint previsions. We show how to derive updated previsions using both natural and regular extension and illustrate the difference between both procedures. Starting from the rationality criteria for conditional previsions – separate and joint coherence –, we derive the generalized Bayes's rule, which provides a practical way of obtaining updated previsions using natural extension. To finish, we show how to obtain joint previsions from a marginal and a conditional prevision using marginal extension and also how to combine a collection of marginals using independent products.

After all this is digested, Chapter 2$_{66}$, Chapter 3$_{94}$, and Appendix G$_{220}$ have become accessible.

0.2.2 *Extreme lower probabilities*

Chapter 2, 'Extreme lower probabilities'$_{66}$, is not about learning from samples. It is about representing the set of all lower probabilities that satisfy a given set of properties, and this for different sets of properties.

This material is included first of all because I have spent a nonnegligible amount of time on it; secondly, it fits in this thesis because it shows for conceptually very simple cases – small, finite possibility spaces and only looking at events, not gambles – how properties function as constraints on the possible forms a lower previsions can adopt. Essentially the same thing is done in the next chapter, but for a case that can hardy be called simple.

Extreme lower probabilities are the things we are looking for in this chapter. They are the extreme points of the polytopes that sets of all lower probabilities satisfying a given set of properties are. The section 'Constraints & vertex enumeration'$_{66}$ shows this and also introduces some useful concepts from polytope theory: (sets of) linear constraints, vertices, and vertex enumeration. This last concept describes the step of going from a set of constraints that define a polytope (the way most properties for lower probabilities are specified) to the defining set of vertices of that polytope (which for us is the set of extreme lower probabilities).

So, once the set of constraints corresponding to some property has been enumerated, going to the corresponding vertices is a solved, but computationally still very intensive problem: one just uses one of the freely available vertex enumeration computer programs. Therefore, in 'Property constraint generation'$_{70}$, we derive manageable sets of constraints for many of the properties that lower probabilities are commonly required to satisfy. These properties are normedness, nonnegativity, additivity, k-monotonicity, avoiding sure loss, coherence, permutation invariance, and maxitivity. We have written a computer program that is capable of generating all the necessary constraints for each of these properties and for all finite possibility space sizes; the only limiting factor is the available computing memory and time.

Successively using our constraint generation program and the available vertex enumeration programs results in a list of vertices that corresponds to the sets of extreme lower probabilities we were looking for. The most instructive of these sets are analyzed in the 'Results'$_{85}$ section; some arguably even more interesting ones are analyzed in Appendix A, the Herbarium$_{194}$. This appendix also contains two technical lemmas used in the derivation of manageable sets of constraints for avoiding sure loss and coherence.

0.2.3 *Inference models*

With Chapter 3, 'Inference models'$_{94}$, the focus is squarely on learning from samples. Samples from a finite set of categories, to be precise. This provides an amply general context for all the important issues in the construction of inference models – i.e., models that allow us to learn from samples – to appear.

The exchangeability assumption studied in the first section, 'Ex-

changeability, sufficient statistics & likelihood functions'$_{95}$, lies at the basis of all the inference models we encounter. Informally, it states that the order in which samples are observed is unimportant. This results in a strong constraint on the model that describes the uncertainty about a finite or an infinite sequence of samples: we obtain two de Finetti-esque representation theorems. This restriction is also reflected in the lower previsions that correspond to the updated uncertainty models obtained after observing a partial sequence of samples: they only depend on a so-called sufficient statistic of the observation through a so-called likelihood function. It is these lower previsions that constitute an inference model; therefore, this inference model is heavily impacted by the exchangeability assumption. At the end of this section, I make the link with classical Bayesian updating and introduce the distinction between predictive inference models (for observables) and parametric inference models (for abstract mathematical objects).

Once we have a good view on the impact of the exchangeability assumption, we move on and use this knowledge to construct two closely related inference models introduced by Walley & Bernard. The first, in 'Predictive inference: representation insensitive prediction'$_{118}$, is the predictive imprecise Dirichlet-multinomial model; the second, in 'Parametric inference: the imprecise Dirichlet model'$_{138}$, is the parametric imprecise Dirichlet model. We actually derive these models from first principles, i.e., from the exchangeability assumption, from a requirement of representation insensitivity – i.e., the choice of the possibility space the observations are embedded in must not matter –, by imposing that immediate predictions are simple in the sense that they are linear-vacuous, and by adding a technical simplification called following the constant hyperparameter path. We also point out that the imprecise Dirichlet model can also be seen as an inference model based on updating a so-called prior set of parametric linear previsions that are conjugate to the likelihood function; this is also the way Walley introduced this model.

We end the chapter with 'Applications'$_{142}$, a section that shows how the imprecise Dirichlet-multinomial model can be put in practice by applying it to game-theoretic learning and the learning of Markov chains.

0.2.4 *Inference models for exponential families*

In Chapter 4, 'Inference models for exponential families'$_{154}$, we continue learning from samples, but now broaden our scope from categorical sampling to sampling models described by exponential family likelihood functions. We base the inference models we propose on updating sets of parametric linear previsions that are conjugate to these likelihoods.

Therefore, in 'Exponential families & friends'$_{155}$, we give an overview of what these exponential families look like and derive corresponding

conjugate parametric and predictive families of previsions. As is also the case for the conjugate parametric and predictive previsions underlying the imprecise Dirichlet and Dirichlet-multinomial models, these previsions can be parameterized using a number of counts and a mean single-sample sufficient statistic. We illustrate this overview and derivation with two examples, one about the well-known normal sampling model and one to show that categorical sampling also fits into this framework. Other examples can be found in Appendix B, the Bestiarium$_{202}$.

In 'Imprecise-probabilistic inference models for exponential families'$_{173}$, we introduce imprecise-probabilistic parametric and predictive inference models by considering prior sets of conjugate parametric and predictive linear previsions with fixed count parameter, but varying mean single-sample sufficient statistic parameter. There was actually some freedom in the choice of the conjugate family; we picked that family for which immediate predictions about the sufficient statistic are simple in the sense that they are (more or less) linear-vacuous. To give some feeling for what these inference models look like, we continue the examples of the foregoing section.

The last section, 'Application: naive credal classification'$_{180}$, gives a description of how the inference models just introduced – in their immediate predictive incarnation – can be used for credal classification. For this, we first describe what credal classification is, then sketch Zaffalon's naive credal classifier for categorical data, and finally generalize this naive classifier to data from exponential family sampling models.

0.2.5 *Lower & upper covariance*

Appendix G, 'Lower & upper covariance'$_{220}$, gives, as can be gathered from the title, a definition of lower and upper covariance and the reasoning leading to this definition. Lower and upper covariances are generalizations to the theory of coherent lower previsions of the classical concept of covariance. The reasoning leading to the definition mimics the one used by Walley for lower and upper variances.

This material has been included because it was my first original result in the theory of imprecise probabilities during my doctoral research.

The appendix's letter is chosen as a tribute to Walley [1991].

0.3 Basic mathematical concepts & notation

We try to follow Boute's [2005] mathematical notation in this thesis because of its cleanliness, formal correctness, and close similarity to most typically found notations. To further improve familiarity, we have consciously made some concessions and modifications, however.

To make sure we can start from common ground, we here give an overview of the basic mathematical concepts and notation used.

0.3.1 *The placeholder, number sets, intervals & extrema*

Let us begin by introducing the notation for nothing: We use the placeholder $\bullet$ to indicate the affix convention of functions or to replace variables a function has not been evaluated in. For example, multiplication uses the infix notation $\bullet\cdot\bullet$, and the relationship between the cosine function and complex exponentials is $2\cdot\cos = \exp(\mathrm{i}\cdot\bullet) + \exp(-\mathrm{i}\cdot\bullet)$. As you can see in the last example's left-hand side, the placeholder is omitted when this can be done without ambiguity.

Common set operators: union $\cup$, intersection $\cap$, difference $\setminus$, and cardinality $|\bullet|$.

The number sets we come in contact with are the reals $\mathbb{R}$, the rationals $\mathbb{Q}$, the integers $\mathbb{Z}$, the nonnegative integers or naturals $\mathbb{N}$, and the Booleans $\mathbb{B} = \{0,1\}$, which are also used as truth values – like de Finetti [1972*a*] does. Not specifically a number set – but still commonly used – is the empty set $\emptyset$.

Closed, open, closed-open, and open-closed intervals of the reals are respectively written $[\bullet,\bullet]$, $]\bullet,\bullet[$, $[\bullet,\bullet[$, and $]\bullet,\bullet]$. Intervals of integers are written $\bullet..\bullet$. For example, $-1..1 = \{-1,0,1\}$. If a (closed) interval's upper bound is (strictly) lower than its lower bound, it corresponds to the empty set $\emptyset$.

When a set $\mathscr{S}$ has sufficient order structure, it can have a minimum $\min\mathscr{S}$, a maximum $\max\mathscr{S}$, an infimum $\inf\mathscr{S}$, and a supremum $\sup\mathscr{S}$. For example, $\sup[0,1[= 1$.

0.3.2 *Bindings & definitions*

Every new mathematical object with an identifier such as x is introduced by binding it to some set $\mathscr{X}$ with $x:\mathscr{X}$. The binding symbol : can be read as 'in'.

Negated proposition: $\neg(\alpha\in\mathbb{Z}) \Leftrightarrow \alpha\notin\mathbb{Z}$.

Optionally, one can add a proposition p that x has to satisfy. We write this more concisely as $x:\mathscr{X}\wedge p$. The filter symbol $\wedge$ can be read as 'such that'. For example, with $\alpha:\mathbb{Q}$, we make α a rational number; the proposition $3\cdot\alpha\in\mathbb{Z}$ further restricts its possible values.

Basic sets are introduced without giving a binding.

A binding of x becomes a definition when the proposition p makes x unique. This typically happens when p is a simple equality $x = e$ (where e is an expression only containing constants); in this case, the set the object belongs to can be unambiguously determined from its defining proposition and the shorthand $x := e$ can be used. So $\alpha := 1/3$ says the same thing as $\alpha:\mathbb{Q}\wedge\alpha = 1/3$.

The power set function $\wp$ returns the set of all subsets of any given set.

Other useful binding shorthands are those that introduce subsets. For example, $A:\subseteq\mathbb{N}$ is equivalent notation for $A:\wp\mathbb{N}$. (The binding symbol : itself can be seen as a less symbol-heavy shorthand for $:\in$.)

In the flow of text, we often replace formal bindings by some equally informative wording. For example, $\beta:\mathbb{N}\wedge\beta/2\in\mathbb{N}$ could be replaced by 'the even natural number β'.

0.3.3 *Functions & abstractions*

Apart from sets and numbers, every mathematical object is looked at as a function (or operator): It is defined by specifying a set as the domain and giving an expression that defines a rule for assigning values to every element of the domain. The set of all these values is the range of the function.

Function sets are defined using the function arrow $\bullet \to \bullet$. For example, $\mathbb{R} \to \mathbb{Z}$ is the set of all integer-valued functions on the real numbers: the domain is $\mathbb{R}$ and the image is $\mathbb{Z}$; the range of a function in such a set need not be the whole image. Bijection sets are introduced with the bijection arrow $\bullet \leftrightarrow \bullet$ In worded bindings, we usually write things like 'the irrational-valued function f on the set of prime numbers'.

We *do not* use parentheses for function application, but only for overriding operator precedence rules and increasing legibility. Examples are (let $f : \mathbb{R} \to \mathbb{R}$ and $x : \mathbb{R}$) $f\,x$, $f(5 \cdot x)$, $f\,x^2$, $(f\,x)^2$, and – more generally – partial application of binary operators; e.g., $(+5)3 = 3 + 5$.

Functions that take the role of coefficients get their argument as a subscript; coefficient sets are defined using superscript notation. For example, $\lambda : \mathbb{R}^{\{a,b,c\}}$ has components λ_a, λ_b, and λ_c.

One way to introduce functions is with a so-called abstraction

$$x : \mathscr{X} \wedge p \,;\, e, \tag{1}$$

which links a rule e to the binding $x : \mathscr{X} \wedge p$ for the dummy variable x. For example, $x : \mathbb{R} \,;\, x^2$ is a parabola function and one of its right inverses is $y : \mathbb{R} \wedge y \geq 0 \,;\, -\sqrt{y}$. For constant functions, there is no need to mention the dummy variable: $(X \,;\, \mu)$ is the μ-valued constant function on X.

The square root function $\sqrt{\bullet} := (\bullet)^{1/2}$ grows a vinculum for grouping; e.g., $\sqrt{1+\pi} = \sqrt{(1+\pi)}$.

An abstraction does not specify an image, but the range is implicitly defined. There are two syntactic variants for special abstractions:

$$x : \mathscr{X} \mid p \text{ stands for } x : \mathscr{X} \wedge p \,;\, x, \text{ and} \tag{2}$$

$$e \mid x : \mathscr{X} \text{ stands for } x : \mathscr{X} \,;\, e. \tag{3}$$

Used in combination with the range function $\{\bullet\}$, which returns the range of a function, this leads to familiar expressions such as $\{\beta^2 \mid \beta : \mathbb{R}\}$ and $\{\alpha : \mathbb{Q} \mid 3 \cdot \alpha \in \mathbb{Z}\}$. When an extremum operator is applied to this last type of set, we also allow the binding to be placed as a subscript; for example, $\min\{\beta^2 \mid \beta : \mathbb{R}\} = \min_{\beta:\mathbb{R}} \beta^2$.

0.3.4 *Predicates & the quantifiers*

Predicates are functions with $\mathbb{B}$ as an image set. A typical example of predicates are the binary relations. We already encountered equality $=$ and belongs to $\in$. Other examples are proportional to $\propto$, smaller than $\leq$ and strictly smaller than $<$, subset of $\subseteq$ and strict subset of $\subset$. Partial

Of some binary relations we also use the negated versions, e.g., inequality $\neq$.

applications of relations are predicates too: (> 0) is the predicate that holds for all strictly positive numbers.

Predicates attached to sets by subscripting restrict the set in question to those elements for which the predicate holds. For example,

$$\mathbb{N} = \{k : \mathbb{Z} \mid k \geq 0\} = \mathbb{Z}_{\geq 0}.$$

Similarly, sets are used as subscripts to functions to restrict or trivially extend their domain. An example of the former technique: the identity map id that returns its argument is typically used after restriction to some domain; e.g., $\mathrm{id}_{\mathbb{R}} \in \mathbb{R} \to \mathbb{R}$. Trivial extension defines a function to be zero in those parts of the new domain it was previously undefined.

The universal quantifier $\forall$ and existential quantifier $\exists$ are predicates of predicates. For example, consider the real-valued polynomial v on $\mathbb{R}$, then $\forall x : \mathbb{R} ; vx = 0$ holds if v is zero everywhere and $\exists x : \mathbb{R} ; vx = 0$ holds if v has real roots somewhere.

Mathematical constants are written upright; e.g., $e^{i \cdot \pi} = -1$.

Quantified abstractions are often nested: $\forall x : \mathbb{R} ; \exists y : \mathbb{R} ; x + y = \pi$, for example. When the expression gets too long, it is structured on multiple lines to improve readability (e.g., $(1.25)_{40}$).

0.3.5 *Pointwise extension*

We implicitly (partially) pointwise extend all common binary arithmetic operations. For example, let α be some real number and f and g real-valued functions on $\mathbb{R}$, then the pointwise extension of addition, multiplication, and exponentiation are illustrated by the following equalities:

$$f + g = x : \mathbb{R} ; fx + gx, \qquad \alpha + g = x : \mathbb{R} ; \alpha + gx,$$
$$f \cdot g = x : \mathbb{R} ; fx \cdot gx, \qquad f^{\alpha} = x : \mathbb{R} ; (fx)^{\alpha}.$$

Apart from exponentiation, the same can be done for appropriate sets; e.g., $\mathbb{Z} = \mathbb{N} - \mathbb{N} = \{m - n \mid m, n : \mathbb{N} \times \mathbb{N}\}$.

We also implicitly (partially) pointwise extend all common binary relations, *but* at the same time add universal quantification. For example, the extension of the 'smaller than'-relation can be illustrated by the following equivalences:

$$f \leq g \Leftrightarrow \forall x : \mathbb{R} ; fx \leq gx, \qquad \alpha \leq g \Leftrightarrow \forall x : \mathbb{R} ; \alpha \leq gx.$$

Returning to the example with the polynomial v near the end of the previous subsection, we can use this pointwise extension to write $v = 0$, to be read as 'v is zero everywhere (in its domain)'.

0.3.6 *Sequences, tuples & elastic operators*

Sequences and tuples are also functions; the former have some denumerably infinite set as their domain – typically $\mathbb{N}$ – and the latter a finite domain – typically $\mathbb{N}_{<n}$, where $n : \mathbb{N}$ is the tuple's length. A tuple is written as a comma-separated list of mathematical objects. A vector is a tuple of

elements of the same – restricted – set. For example, let $\rho := 1/3, 2/3$, then $\rho 0 = (1/3, 2/3)0 = 1/3$ and, using the range function, $\{\rho\} = \{1/3, 2/3\}$. This example also shows that we need a singleton function $\imath$, whose value in any argument is the singleton set containing that argument; otherwise, we cannot write down the set $\imath\rho$ that only contains the vector ρ.

Elastic operators are functions that can take sequences, tuples, and sets of arbitrary size as arguments. Typical elastic operators are sums $\sum$, products $\prod$, unions $\bigcup$, and intersections $\bigcap$. For example, when working with vectors or sets of sets, let $m := 1, 2$ and $J := \{\emptyset, \{2,3\}, \{2,5\}\}$, then

$$\sum m = \sum(1,2) = 1+2, \qquad \bigcup J = \bigcup\{\emptyset, \{2,3\}, \{2,5\}\} = \{2,3,5\},$$
$$\prod m = \prod(1,2) = 1\cdot 2, \qquad \bigcap J = \bigcap\{\emptyset, \{2,3\}, \{2,5\}\} = \emptyset.$$

With abstractions, the binding can be moved to subscript location, so let $f := k : \mathbb{Z} ; \frac{3}{7}\cdot 2^{-|k|}$, then

$$\sum f + \min\{f^2\} = \sum_{k:\mathbb{Z}} \tfrac{3}{7}\cdot 2^{-|k|} + \min_{k:\mathbb{Z}} \tfrac{9}{49}\cdot 4^{-|k|},$$

To make the above unambiguous: subscripted elastic and extremum operators have higher precedence than addition, but lower precedence than multiplication and function application.

Set and function restriction can be used to compactly write down some elastic operations; e.g.,

$$\bigcap J_{\neq\emptyset} = \{2,3\} \cap \{2,5\} = \imath 2 \quad \text{and} \quad \sum f = 2\cdot \sum f_{\mathbb{N}} - \tfrac{3}{7}.$$

Note that the quantifiers $\forall$ and $\exists$ can be seen as the elastic operators respectively associated with conjunction $\wedge$ and disjunction $\vee$.

0.3.7 *Integrals*

Notation for integrals $\int$, the most common linear functions of functions with a nondiscrete domain, follows classical conventions. For example, let $g := x : \mathbb{R} ; \exp(-x^2)$, then

$$\int_{[-1,1]} g = \int_{[-1,1]} g x \,\mathrm{d}x = \int_{[-1,1]} \exp(-x^2)\,\mathrm{d}x = \int_{-1}^{1} \exp(-x^2)\,\mathrm{d}x,$$

denotes its Lebesgue integral of g over $[-1,1]$. (We do not mention Lebesgue measure explicitly because we only use this one type of integral.)

MODELING UNCERTAINTY
The theory of coherent lower previsions

reason, *v.i.* To weigh probabilities in the scales of desire.
Bierce [1911]

Models for uncertainty can be divided in two types, depending on their goal. The first are the descriptive models, which try to capture how people deal with uncertainty in practice. The second are the normative models, which prescribe how people *should* deal with uncertainty in practice.

We are interested in normative models: we wish to provide people with tools that help them solve problems involving uncertainty in a reasonable way. What *reasonable* means, must be agreed upon when choosing an uncertainty model.

In this chapter, we work towards a very expressive theory for modeling uncertainty: the theory of coherent lower previsions [Walley 1991]. It is expressive in the sense that it encompasses classical probability theories and various other, nonclassical ones. Furthermore, its mathematical language can be used for various application domains and with different interpretations of the nature of uncertainty. In this chapter, the basic definitions and properties of this theory are presented and illustrated. Only those parts are elaborated that are needed further on, or that have received special attention during my research.

A general remark about modeling applies here too: We do not strive to obtain universal generality; there are always problems for which the theory cannot provide a fitting model. Considering our needs (uncertainty models that can form a basis for learning from samples), we feel that the theory for modeling uncertainty we sketch and illustrate in this chapter strikes a nice balance between applicability and complexity.

In §1.1, we give an overview of the ideas underlying the mathematical formalization of the theory of coherent lower previsions. After that, we explore the mathematical consequences of the rationality criteria we follow (in §1.2$_{34}$) and sketch how the encountered uncertainty models can be restricted, transformed, and combined (§1.3$_{52}$). Those familiar with the theory of coherent lower previsions should at least skim this chapter to get acquainted with the notations and conventions we introduce.

1.1 Formalizing uncertainty

A prime aspect of modeling uncertainty is formalizing it in such a way that mathematics can be used. This is done by making mathematical objects out of the things we are uncertain about and placing them on some scale, for example by attaching a value to them.

1.1.1 *Events, the possibility space & its subsets*

A most basic thing to be uncertain about is the occurrence of an event, usually seen as a logical statement that can (in principle) be verified to be either true or false. For example, the famous event 'the sun will rise tomorrow' is commonly held to be true; the interested reader can verify this by waiting not more than one day.

We use $\mathscr{E}$ as notation for a generic set of events.

In a lot of problems, a possibility space Ω can be identified as part of the modeling effort: It is a sufficiently exhaustive set of possible but mutually exclusive so-called elementary events. Multiple (possibly interrelated) possibility spaces may be identified for the same problem and used consecutively or concurrently when working on the problem. To select ingredients for a meal, for example, it is advisable to use a possibility space that at least distinguishes between what is and what is not edible. Less hungry or more experienced chefs end up employing more refined possibility spaces.

The elementary events can be represented by the elements of Ω and any event by a subset of Ω, or equivalently, as an element of its power set $\wp\Omega$. Both can also be represented by predicates in $\Omega \to \mathbb{B}$: The predicate that corresponds to some subset A of Ω is defined as $I^A := (\in A)$ and is called the indicator of A. Similarly, the indicator for an elementary event $\omega : \Omega$ is defined as $I^\omega := (= \omega)$.

The terminology used depends on the kind of problem under consideration: Quite often, the elementary events differ only in some defining characteristic – the state – of the universe of discourse they refer to. So, elementary events and states can play the same role. For example, the state space $\{\text{red}, \text{green}, \text{blue}\}$ can refer to the three types of color-sensitive cones in the typical human eye or to M&M's. Also, whenever the elementary events or the states can be seen as samples generated by some process, the term sample space is used as well.

1.1.2 *Random variables & bounded functions*

Next to the occurrence of an event, another basic thing to be uncertain about is the value of some variable, usually seen as a phrase that implies some value that can (in principle) be determined. For example, forecasters tend to be notoriously unsure about 'the amount of rainfall tomorrow', a quantity which any disinterested bucket will slavishly mea-

sure. The word random is classically added to distinguish them from variables whose value is certain, such as 'the author of this thesis'.

We use $\mathcal{K}$ as generic notation for a set of random variables.

The range of values a random variable may assume is usually considered to be known. In this thesis, only bounded subsets of the reals occur, except when random variables are used to distinguish subsequent observations belonging to the same sample space. The element of the range that is actually observed is called the realization of the random variable. Events can be seen as boolean-valued random variables.

Again, in case some possibility space Ω can be identified, random variables can be represented by functions on Ω. The set of representations as functions of all bounded real random variables is written $\mathscr{L}_\Omega$. The set of all indicators, representations of events, is defined by

$$\mathscr{I}_\Omega := \Omega \to \mathbb{B}. \tag{1.1}$$

1.1.3 *Two basic models: preference orders & desirability*

Once we have modeled what we are uncertain about, we can model the uncertainty itself. To wit, once we have defined some set of events or some set of random variables, we can model our uncertainty by adding structure to the set, such as an order, or by adding labels to its elements.

Consider two distinct events A and B of some set $\mathscr{E}$ of events. Expressing our uncertainty can be as simple as stating that A is more probable than B, which adds some order to $\mathscr{E}$. Similarly, some set $\mathcal{K}$ of random variables corresponding to the different types of tickets for some lottery can be given an order by specifying preferences between tickets on the basis of the expected winnings and losses.

Uncertainty about the outcome of a lottery can also be modeled by calling those lottery tickets for which we expect winnings 'desirable'; the other tickets are then non-'desirable'. (We could make a more expressive model by labeling as 'undesirable' those non-'desirable' tickets for which we expect losses.)

1.1.4 *Probabilities, expectations & previsions*

Another way of labeling events and random variables is by attaching one or more real values to them, instead of one out of a discrete set of labels as with desirability for random variables. Attaching more than one value could be useful, for example, when we do not know whether the lottery numbers are drawn with or without replacement; one just attaches a value for each of these two cases.

A value attached to some event to indicate the likelihood of its occurrence is usually called a probability of this event. Given a set of events $\mathscr{E}$, a probability can be seen as a function in $\mathscr{E} \to \mathbb{R}$. Typical examples of

probabilities are betting rates and the relative frequency of different causes of death.

Similarly, a value attached to some random variable that gives some kind of estimate for its expected realization, is called an expectation of this random variable. Given a set of random variables $\mathcal{K}$, an expectation is seen as a function in $\mathcal{K} \to \mathbb{R}$. Examples are acceptable buying or selling prices for lottery tickets and the design height of dikes.

The term prevision is usually – and in this thesis – used as a synonym for expectation; we, moreover, let it take up the role of both a probability and an expectation. As events (in $\mathscr{E}$) can be seen as special random variables (in $\mathcal{K}$), a prevision is a function in $\mathcal{K} \cup \mathscr{E} \to \mathbb{R}$ whose value in some event is defined by its value in the corresponding random variable.

De Finetti [1972*a*] and Walley [1991] also 'use the same symbol' for both probability and expectation.

Whenever a possibility space Ω can be identified, the prevision of some subset or element of Ω is defined by the prevision in the corresponding indicator. Given a set of functions $\mathcal{K} :\subseteq \mathscr{L}_\Omega$, we can enlarge it with the sets and elements corresponding to the included indicators to define

$$\mathcal{K}^\star := \mathcal{K} \cup \{B :\subseteq \Omega \mid I^B \in \mathcal{K}\} \cup \{\omega : \Omega \mid I^\omega \in \mathcal{K}\}. \tag{1.2}$$

For example, ${\mathscr{L}_\Omega}^\star = \mathscr{L}_\Omega \cup \wp\Omega \cup \Omega$ and ${\mathscr{I}_\Omega}^\star = \mathscr{I}_\Omega \cup \wp\Omega \cup \Omega$.

The predicate $\operatorname{ind} : (\mathcal{K}^\star \to \mathbb{R}) \to \mathbb{B}$ that checks whether the prevision of some indicator is the prevision of the corresponding subset (or element) of Ω, is defined by (let $P : \mathcal{K}^\star \to \mathbb{R}$)

$$\operatorname{ind} P \Leftrightarrow \begin{cases} \forall B : \wp\Omega \cap \mathcal{K}^\star ; PB = PI^B \\ \forall \omega : \Omega \cap \mathcal{K}^\star ; P\omega = PI^\omega. \end{cases} \tag{1.3}$$

This predicate allows us to introduce the often-used function that generates the set of all previsions from a set of functions; to wit,

$$\mathscr{P}\mathcal{K} := (\mathcal{K}^\star \to \mathbb{R})_{\operatorname{ind}} \tag{1.4}$$

defines the set of all previsions on $\mathcal{K}$.

We will not encounter ind again, but it provides a first example of what we do multiple times later on: use predicates to introduce properties and give definitions.

1.1.5 *Meaning & measurement*

Above, we have used the words 'probable', 'expect', and 'likelihood' informally. This is sufficient for a first introduction of some concepts, but it is insufficient for building normative uncertainty models. We need to give a meaning (or interpretation) to the structure added to the sets of events and random variables or to the labels added to its elements. This meaning

(i) guides the formulation of the calculation rules,

(ii) guides the specification of what constitutes a reasonable uncertainty model, and
(iii) allows us to provide interpretable results.

We also need compatible measurement approaches for obtaining an uncertainty model, i.e., such a structure or such labels.

There is no uniquely accepted interpretation attached to uncertainty models. Nor should there be: different application domains have different requirements. Competing with the vast literature on this subject is not an aim of this thesis. Let it suffice to say that one interpretation is used, but our results might still be compatible with others.

Fine [1973] gives a nice overview of possible interpretations.

This interpretation is based on the one advanced by Walley [1991]. My reading of it, influenced by my supervisor, other colleagues & de Finetti [1972*b*] and colored by personal nuance, is the following:

- Uncertainty is all in the mind: We model an abstract degree of belief of the concerned parties, which is based on the knowledge about what is modeled; therefore, the uncertainty model is (inter)subjective and epistemic. Things regarded as empirical fact do not have a special status, but are directly assimilated; there is no need for a separate aleatory uncertainty model.
- Uncertainty is handled with reason: Rationality criteria that define what constitutes a reasonable model must be agreed upon by the concerned parties (both universal criteria, e.g., avoiding sure loss, and problem-specific assumptions, e.g., exchangeability).
- Models are evidence-based: Influences that go beyond the rationality criteria should be based on evidence (e.g., the composition of a sample). Assumptions based on mathematical convenience and arbitrary choice inevitably also play a role, but should of course be identified as such.
- Behavior is revealing and prescribeable: The abstract degree of belief is reflected in the behavior of the concerned parties (as a whole). Furthermore, behavior implied by the uncertainty model can be followed (and *should* be, if the model is taken seriously).

The second and third point could have been omitted, as they are indirectly implied by the first. Their importance lies in the creation of a mental framework: They concern the partitioning of beliefs into particular types (belief about what is reasonable, about what is evidence, and about acceptable simplifications). These types of belief are represented and used in different ways, with the common goal of solving problems involving uncertainty in a structured way. In the end, the concerned parties must judge whether the theory and the resulting models are acceptable. For them to be able to do this, *all assumptions made must be explicitly stated.*

The measurement approaches advanced in this thesis are all some form of assessment, i.e., construction from samples using a partly arbi-

trary inference model. The arbitrariness could be reduced by additionally using a second measurement approach, elicitation, i.e., observing behavior (such as choices and assertions made).

1.1.6 *Betting behavior & utility*

The types of behavior influenced by uncertainty vary widely, but this is of little concern to us, as creating a general descriptive model is not our aim. We limit the scope of our normative model to a manageable size by only considering behavior related to dealing with gambles (i.e., accepting, rejecting, buying, and selling them).

The prototypical example is modeling the uncertainty involved in a lottery using different lottery tickets: As we have already seen, preferences between them can be declared, they can be labeled as desirable or not, and acceptable buying and selling prices for them can be specified. The problems falling within our scope are those that can be (abstractly) reframed as some kind of lottery using some kind of gambles.

Betting behavior and linear precise utility are also important building blocks in Smith's [1961] seminal sketch of a theory of lower and upper probabilities.

We assume that the positive or negative reward of a gamble is expressed in some precise and linear utility. Roughly, this means that, for the concerned parties, there is no doubt about the value of some reward and that getting the same reward twice is twice as useful as getting it once. This is a simplifying assumption made to reduce model complexity; it limits the use of the uncertainty model to problems for which linear precise utility is an acceptable approximation.

1.1.7 *Lower and upper probabilities & previsions*

In this thesis, the uncertainty models we mainly use are previsions and probabilities. We now introduce these two concepts in terms of betting behavior.

The betting rate is the proportion of the amount that can be lost to the amount that can be won (the stake).

When using events, the betting framework can be used if their occurrence can be verified at some point. Uncertainty models are then defined by thinking in terms of betting on or against the occurrence of an event. The supremum acceptable betting rate for an event is called the *lower* probability of this event. The *upper* probability of an event is one minus the supremum acceptable betting rate against this event. Higher supremum betting rates on an event imply more of a commitment.

When using random variables, the framework can be used if the actual realization of the random variables can be observed. Within the framework, random variables are seen as gambles, and their realization as the reward (either winnings or losses) for the owner. The supremum acceptable buying price for obtaining a gamble is called its *lower* prevision, the infimum acceptable selling price its *upper* prevision. Higher supremum buying prices and lower infimum selling prices for a gamble imply more of a commitment.

Recall that events can be seen as a special type of random variables; thus bets can be seen as a special type of gambles. Therefore, probabilities are seen as restricted previsions.

Following Walley [1991], random variables that are bounded real-valued functions whose values are interpreted as rewards are from now on called gambles, a constant reminder of the framework we are working in. Consider a possibility space Ω. As the theory will from this point onward be worked out in terms of gambles, the notation $\mathscr{K}$ will from now on be used to refer to a set of gambles only. Similarly, $\mathscr{L}_\Omega$ will be used to refer to the set of all gambles on Ω only: let bnd be the boundedness predicate, then

Illustrating the boundedness predicate: $\operatorname{bnd}\sin = 1$, $\operatorname{bnd}\tan = 0$.

$$\mathscr{L}_\Omega := (\Omega \to \mathbb{R})_{\operatorname{bnd}}, \tag{1.5}$$

It is a linear space for pointwise addition and (real) scalar multiplication; we endow it with the supremum-norm topology (also called the topology of uniform convergence).

Walley [1991, §3.7$_{150}$ and §3.8$_{158}$] elaborates on the relationships between previsions, desirability, and preference orders.

The uncertainty models we specify in terms of previsions of gambles can be equivalently specified using preference orders of gambles or as sets of desirable gambles. The latter type of model is in fact used in the next section as the basis for the definition for previsions and their rationality criteria.

1.2 RATIONALITY & ITS CONSEQUENCES

This section is based mainly on Walley's [1991] work, and through him, on Williams's [1974].

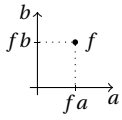
Until now, we have only stressed the importance of rationality criteria, but stayed silent about the actual criteria we impose. The values of previsions are as of yet unrestricted. I can still unabashedly state that my lower and upper probability for a white Christmas are respectively 3 and -1. Such nonsense, you must agree, has to stop here, in this section.

1.2.1 *Desirability*

Apart from the status of the zero gamble, Walley [1991, §F$_{614}$] calls what we consider here 'real desirability'.

We first formally introduce the uncertainty model using sets of desirable gambles and the associated rationality criteria. This serves as a basis for the definition of an almost equivalent uncertainty model consisting of lower and upper previsions and their rationality criteria. We take this route, because the rationality criteria of the former model are easier to formulate and justify.

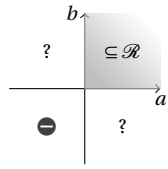
Illustration with $\Omega := \{a, b\}$:



We consider a possibility space Ω and the corresponding set $\mathscr{L}_\Omega$ of all gambles. We generically denote a set of desirable gambles by $\mathscr{R} :\subseteq \mathscr{L}_\Omega$. When is a gamble f on Ω desirable or not? Informally, it is desirable when we expect it to bring us positive winnings; it is not desirable if this is not the case.

It would therefore be very reasonable to consider gambles that can yield a positive reward without ever resulting in a negative reward to

be desirable; similarly, it is reasonable to consider gambles that might yield a negative reward without ever giving a positive reward not to be desirable. This leads to the first two rationality criteria:



$$\text{Accepting partial gains:} \quad \begin{cases} \sup\{f\} > 0 \\ \inf\{f\} \geq 0 \end{cases} \Rightarrow f \in \mathscr{R}, \tag{1.6}$$

$$\text{Avoiding partial loss:} \quad \begin{cases} \inf\{f\} < 0 \\ \sup\{f\} \leq 0 \end{cases} \Rightarrow f \notin \mathscr{R}. \tag{1.7}$$

If, in these two criteria, we replace the nonstrict inequalities by strict ones, we get the criteria for accepting *sure* gains and avoiding *sure* loss.

The criteria above hold independently of the uncertainty that we want to model. Any other gamble might or might not be desirable. However, the fact that we use precise, linear utility does imply two more rationality criteria. First, if a gamble is desirable, i.e., is expected to give a positive reward, then multiples and fractions of this gamble are also desirable. Secondly, if two gambles are expected to give a positive reward, then their sum is also desirable. So our assumptions about the utility used results in the following requirements: (let $f, g : (\mathscr{L}_\Omega)^2$ and $\lambda : \mathbb{R}_{>0}$)

$$\text{Positive scaling:} \quad f \in \mathscr{R} \;\Rightarrow\; \lambda \cdot f \in \mathscr{R}, \tag{1.8}$$

$$\text{Addition:} \quad f, g \in \mathscr{R}^2 \;\Rightarrow\; f + g \in \mathscr{R}. \tag{1.9}$$

A coherent set of desirable gambles $\mathscr{R} :\subset \mathscr{L}_\Omega$ is one that satisfies all four rationality criteria (1.6)–(1.9); it is a convex cone that contains all nonnegative, nonzero gambles.

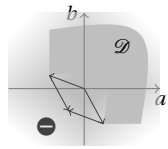
Any partially specified set of desirable gambles $\mathscr{D} :\subseteq \mathscr{L}_\Omega$ can be extended by using (1.6) – i.e., making sure $(\mathscr{L}_\Omega)_{\geq 0 \wedge \neq 0}$ is included – and (1.8)–(1.9) – i.e., taking the positive linear hull. This so-called procedure of natural extension can be seen as a form of deductive reasoning: from an explicitly given set of desirable gambles, we can use our rationality criteria to deduce which other gambles are implicitly desirable. It results in a set of desirable gambles

Introducing the finite subset relation: $\{2,3,5\} \Subset \mathbb{N}$, $\emptyset \Subset \mathscr{L}_\Omega$.

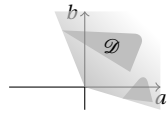
$$\mathscr{R}_\mathscr{D} := \left\{ \sum_{f:\mathscr{D}'} \lambda_f \cdot f \;\middle|\; \mathscr{D}' :\Subset \mathscr{D} \cup (\mathscr{L}_\Omega)_{\geq 0 \wedge \neq 0} \wedge \mathscr{D}' \neq \emptyset \,;\, \lambda : (\mathbb{R}_{>0})^{\mathscr{D}'} \right\}, \tag{1.10}$$

which is called the natural extension of $\mathscr{D}$.

Any set of desirable gambles $\mathscr{D} :\subseteq \mathscr{L}_\Omega$ with a natural extension that contains gambles violating (1.7) incurs a partial loss; such $\mathscr{D}$ are unreasonable uncertainty models and therefore to be avoided. Whenever some gamble in the natural extension leads to a sure loss – (1.7), but with strict inequalities –, the natural extension is even the whole of $\mathscr{L}_\Omega$!

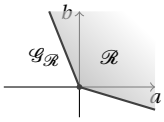


However, any set of desirable gambles $\mathscr{D} :\subseteq \mathscr{L}_\Omega$ for which all gambles in its natural extension avoid a partial loss – i.e., satisfy (1.7) – has a coherent natural extension.



The part of the border of a set of desirable gambles consisting of the gambles with pointwise lowest rewards is the corresponding set of marginally desirable gambles

$$\begin{aligned}\mathscr{G}_{\mathscr{D}} &:= \{f - \sup\{\alpha : \mathbb{R} \mid f - \alpha \in \mathscr{D}\} \mid f : \mathscr{D}\} \\ &= \{f - \sup\{\alpha : \mathbb{R}_{>0} \mid f - \alpha \in \mathscr{D}\} \mid f : \mathscr{D}\}. \end{aligned} \tag{1.11}$$



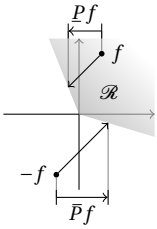
One could call gambles in $\mathscr{R} \cup \mathscr{G}_{\mathscr{R}}$ almost desirable and those in $\mathscr{R} \setminus \mathscr{G}_{\mathscr{R}}$ strictly desirable [cfr. Walley 1991, §3.7$_{150}$].

Note that a marginally desirable gamble is not necessarily desirable. The zero gamble $(\Omega\,;0)$ is always marginally desirable for any coherent set of desirable gambles.

In the next few subsections, we start from what we have learned here to formally introduce lower and upper previsions and their rationality criteria.

1.2.2 *From desirable gambles to lower and upper previsions*

We have already mentioned the interpretation we give to lower previsions $\underline{P} : \mathscr{P}\mathscr{L}_\Omega$ and upper previsions $\overline{P} : \mathscr{P}\mathscr{L}_\Omega$ in §1.1.7$_{33}$: The lower prevision $\underline{P}f$ of a gamble f on Ω is its (finite) supremum acceptable buying price and the upper prevision $\overline{P}f$ of the gamble f its (finite) infimum acceptable selling price.



A buying price $\alpha : \mathbb{R}$ for the gamble f is considered acceptable when the gamble resulting from the transaction $f - \alpha$ is desirable; similarly, a selling price $\beta : \mathbb{R}$ is considered acceptable when the gamble resulting from the transaction $\beta - f$ is desirable. So consider a set of desirable gambles $\mathscr{R} :\subseteq \mathscr{L}_\Omega$; it allows us to write down the following definitions:

$$\underline{P}f := \sup\{\alpha : \mathbb{R} \mid f - \alpha \in \mathscr{R}\}, \tag{1.12}$$

$$\overline{P}f := \inf\{\beta : \mathbb{R} \mid \beta - f \in \mathscr{R}\}. \tag{1.13}$$

Due to the use of suprema and infima in these definitions, the resulting lower and upper prevision is unaffected by whether marginally desirable gambles are desirable or not. So there is a loss of information when going from sets of desirable gambles to previsions. Note that to ensure finiteness of the prices, $\mathscr{R}$ must have a natural extension that avoids sure loss.

We see from (1.12) and (1.13) that lower and upper previsions are related; as an immediate consequence of their definition they satisfy

$$\text{Conjugacy:} \quad \underline{P}(-f) = -\overline{P}f. \tag{1.14}$$

Linearity of utility is at the basis of this relation. The upper prevision of every gamble can be expressed using a lower prevision. Therefore, we can limit ourselves to working out the theory in terms of lower previsions only; we do use upper previsions when convenient notation-wise or interpretation-wise.

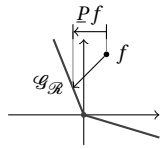
The conjugacy relation for probabilities differs slightly. Consider an event A of Ω; whenever $\underline{P}(1+\bullet) = 1+\underline{P}\bullet$ holds, we can write

$$\text{Conjugacy for probabilities:} \quad \underline{P}(\Omega \setminus A) = 1 - \overline{P}A. \tag{1.15}$$

Imposing conjugacy still does not invalidate the nonsense uttered in this section's introduction. For one thing, requiring a lower prevision to be dominated by its conjugate – i.e., $\underline{P} \leq \overline{P}$ – seems reasonable. This is a consequence of avoiding sure loss, the subject of §1.2.4[39].

But before we can get going with that, we first need a suitable (sort of) converse to (1.12): how to go from a lower prevision to desirable gambles? Let us rewrite (1.12) to see what options we have: ($\mathscr{G}_{\mathscr{R}} := \mathscr{L}_\Omega$ is the set of marginally desirable gambles corresponding to $\mathscr{R}$)

$$\underline{P}f := \sup\{\alpha : \mathbb{R} \mid \exists g : \mathscr{R};\ f - \alpha = g\} \tag{1.16}$$
$$= \sup\{\alpha : \mathbb{R} \mid \exists g : \mathscr{R};\ f - \alpha \geq g\} \tag{1.17}$$
$$= \sup\{\alpha : \mathbb{R} \mid \exists g : \mathscr{G}_{\mathscr{R}};\ f - \alpha \geq g\}, \tag{1.18}$$

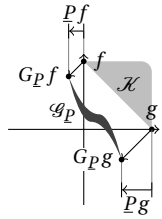


where the respective steps are possible due to the use of suprema. We see that the resulting lower prevision is completely determined by the set of marginally desirable gambles, ignoring the details of the actual borderline behavior.

Now consider a lower prevision $\underline{P}$ on $\mathscr{K}$, where $\mathscr{K} := \mathscr{L}_\Omega$ is the set of gambles for which (finite) supremum acceptable buying prices have been given (a lower prevision does not need to be specified for all gambles). We can then derive the corresponding set of marginally desirable gambles as follows: (cf. (1.11) and (1.12))

If $\underline{P} : \mathscr{P}\mathscr{K}$, then $\overline{P}$ is defined on $-\mathscr{K} = \{-f \mid f : \mathscr{K}\}$.

$$\mathscr{G}_{\underline{P}} := \{f - \underline{P}f \mid f : \mathscr{K}\}. \tag{1.19}$$



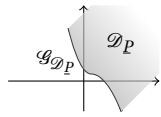
As seen just above, but also in the second paragraph of this subsection, it is useful to think in terms of transactions: buying or selling a gamble for a certain price. Transactions generally play an important role. Therefore, we introduce the so-called marginal gamble: the net result of buying a gamble for its lower prevision. Given a lower prevision $\underline{P}$ on $\mathscr{K}$, the marginal gamble of a gamble is generated by the function $G_{\underline{P}} : \mathscr{K} \to \mathscr{L}_\Omega$, defined for any gamble f in $\mathscr{K}$ by

$$G_{\underline{P}}f = f - \underline{P}f. \tag{1.20}$$

We can see that marginal gambles are marginally desirable by definitions (1.11) and (1.12).

Finally, as the set of desirable gambles corresponding to the lower prevision $\underline{P}$, we take the set

$$\mathscr{D}_{\underline{P}} := \mathscr{G}_{\underline{P}} + \mathbb{R}_{>0} = \{g + \alpha \mid g : \mathscr{G}_{\underline{P}};\ \alpha : \mathbb{R}_{>0}\} \tag{1.21}$$



of gambles consisting of a marginally desirable gamble plus a strictly positive amount of utility to create an acceptable deal. Even though it can

Walley [1991, §3.8.1$_{156}$] seems to consider correspondences between these two types of sets in a more restricted context.

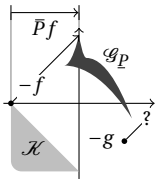
happen that $\mathscr{G}_{\underline{P}} \neq \mathscr{G}_{\mathscr{D}_{\underline{P}}} = \mathscr{G}_{\underline{P}} \setminus \mathscr{D}_{\underline{P}}$, this way of going from marginal gambles to desirable gambles can still be seen as a sort of inverse of (1.11)$_{36}$. We do not claim this to be the only reasonable way to go from a lower prevision via marginally desirable gambles to desirable gambles; for example, another option is $\mathscr{G}_{\underline{P}} + \mathbb{R}_{>0} + (\mathscr{L}_\Omega)_{\geq 0}$, or even $\mathscr{G}_{\underline{P}} + (\mathscr{L}_\Omega)_{>0}$ for finite possibility spaces Ω.

In this and the previous subsection, we have established enough links between sets of desirable gambles and lower previsions to translate the concepts of natural extension, avoiding sure loss, and coherence from the former context to the latter.

1.2.3 *Natural extension*

The main reference for this subsection is Walley [1991, §3.1$_{121}$].

Consider a set of gambles $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ and a lower prevision $\underline{P}$ on $\mathscr{K}$. We are going to look at what our current judgements – formalized as $\underline{P}$ – tell us about the lower prevision of any gamble f in $\mathscr{L}_\Omega$, i.e., about the supremum price we are implicitly willing pay for it. When $f \in \mathscr{K}$, we *correct* our current judgements to make them internally consistent; when $f \notin \mathscr{K}$, we *extend* our judgements. We use the latter term to generically refer to both cases.



The procedure of natural extension for lower previsions is derived by

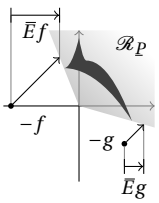
(i) going from the given lower prevision $\underline{P}$ to the corresponding set of marginally desirable gambles $\mathscr{G}_{\underline{P}}$ using (1.19)$_\cap$,

(ii) using this set and (1.21)$_\cap$ to obtain the set of desirable gambles $\mathscr{D}_{\underline{P}}$ that correspond to $\underline{P}$,

(iii) applying the procedure of natural extension for sets of desirable gambles (1.10)$_{35}$ to this set $\mathscr{D}_{\underline{P}}$ to obtain its natural extension

$$\mathscr{R}_{\underline{P}} := \mathscr{R}_{\mathscr{D}_{\underline{P}}}, \tag{1.22}$$

and finally,



(iv) using definition (1.17)$_\cap$ to obtain the lower prevision corresponding to the set of desirable gambles $\mathscr{R}_{\underline{P}}$, resulting in a lower prevision $\underline{E} : \mathscr{P}\mathscr{L}_\Omega$ that is then called the natural extension of $\underline{P}$. In the next subsection, we derive the condition necessary for (1.17)$_\cap$, which ensures that $\mathscr{R}_{\underline{P}}$ avoids sure loss.

We now derive the formula for calculating the supremum acceptable buying price for any gamble f on Ω as it is implied by the given lower prevision $\underline{P}$ and our rationality criteria (1.6)–(1.9)$_{35}$. The equations mentioned in the enumeration above are used in reverse order, the first two are (1.17)$_\cap$ and (1.10)$_{35}$, respectively:

To improve readability, we fix the horizontal location of formula elements; gray lines fill empty spaces.

$$\begin{aligned}\underline{E}f &:= \sup\{\alpha : \mathbb{R} \,|\, \exists g : \mathscr{R}_{\underline{P}};\ \text{——————} \text{——————} f - \alpha \geq g\} \\ &= \sup\{\alpha : \mathbb{R} \,|\, \exists \mathscr{D} :\subseteq \mathscr{D}_{\underline{P}} \cup (\mathscr{L}_\Omega)_{\geq 0} \wedge \neq 0 \wedge \mathscr{D} \neq \emptyset; \\ &\qquad\qquad \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; f - \alpha \geq \textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\}\end{aligned}$$

the use of the supremum allows us to drop the '$\neq 0$'-restriction on $\mathscr{L}_\Omega$, which in turn makes the restriction $\mathscr{D} \neq \emptyset$ moot, so

$$= \sup\{\alpha : \mathbb{R} \mid \exists \mathscr{D} :\subseteq \mathscr{D}_{\underline{P}} \cup (\mathscr{L}_\Omega)_{\geq 0}; \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; f - \alpha \geq \textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\}$$

now we can ignore gambles in $(\mathscr{L}_\Omega)_{\geq 0}$ because the zero gamble is included with the case $\mathscr{D} := \emptyset$:

$$= \sup\{\alpha : \mathbb{R} \mid \exists \mathscr{D} :\subseteq \mathscr{D}_{\underline{P}}; \text{————} \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; f - \alpha \geq \textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\}$$

because of $(1.21)_{37}$ and the use of the supremum, we can replace $\mathscr{D}_{\underline{P}}$ by $\mathscr{G}_{\underline{P}}$; subsequently, we can apply $(1.19)_{37}$:

$$= \sup\{\alpha : \mathbb{R} \mid \exists \mathscr{D} :\subseteq \mathscr{G}_{\underline{P}}; \text{————} \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; f - \alpha \geq \textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\}$$
$$= \sup\{\alpha : \mathbb{R} \mid \exists \mathscr{D} :\subseteq \mathscr{K}; \text{————} \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; f - \alpha \geq \textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot G_{\underline{P}} g\}.$$

We use the last formula – rewritten and reformatted a bit – to introduce the least committal extension function, which maps a lower prevision $\underline{P}$ on $\mathscr{K}$ on whatever domain $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ to its natural extension, so for any gamble f in $\mathscr{L}_\Omega$ it is defined by

$$\operatorname{lce}_{\underline{P}} f := \sup\Big\{\alpha : \mathbb{R} \Big| \begin{array}{l} \exists \mathscr{N} :\subseteq \mathscr{K}; \\ \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{N}}; \\ \qquad f - \alpha \geq \textstyle\sum_{g:\mathscr{N}} \lambda_g \cdot G_{\underline{P}} g \end{array}\Big\}. \tag{1.23}$$

Approximating f by affine positive combinations of gambles gives an equivalent formula [Walley 1991, §3.1.3$_{124}$].

Its name's origin is clarified later on in §1.2.7$_{46}$, where we put it next to other possible extensions. Consider some gamble f in $\mathscr{K}$, by taking $\mathscr{N} := \imath f$ and $\lambda_f := 1$, we see that only upward, nonnegative, corrections are possible.

We now have a way to extend any set of prices (specified by $\underline{P}$ on $\mathscr{K}$) to prices for all gambles. These latter prices are only reasonable whenever $\underline{P}$ is a reasonable uncertainty model; what this means is what we are going to look at next.

1.2.4 *Avoiding sure loss*

In §1.2.1$_{34}$, we saw that sets of desirable gambles that incur a partial loss are unreasonable uncertainty models. This corresponded to their natural extension containing nonzero gambles in the nonpositive orthant $(\mathscr{L}_\Omega)_{\leq 0}$. The interior of this orthant contains gambles leading to a sure loss, its border contains gambles leading to a partial loss that is not sure.

The main reference for this subsection is Walley [1991, §2.4$_{67}$].

For a lower prevision $\underline{P}$ on $\mathscr{K}$, the corresponding sets of (marginally) desirable gambles $\mathscr{G}_{\underline{P}}$, $\mathscr{D}_{\underline{P}}$, and $\mathscr{R}_{\underline{P}}$ defined by $(1.19)_{37}$, $(1.21)_{37}$, and (1.22) are the ones to look at. Formally, they incur a partial loss or a sure loss, if

respectively (cf. $(1.7)_{35}$)

$$\exists g : \mathscr{R}_{\underline{P}} ; \inf\{g\} < 0 \wedge \sup\{g\} \leq 0, \quad \text{or} \quad \exists g : \mathscr{R}_{\underline{P}} ; \sup\{g\} < 0. \tag{1.24}$$

By taking their converse and doing some rewriting and reformatting (in the spirit of the derivation of $(1.23)_{\frown}$, which we shall from now on call back-expanding), we can obtain expressions for avoiding sure or partial loss more explicitly in terms of $\underline{P}$ and $\mathscr{K}$. One step in the derivation of the criterion for avoiding partial loss is important; to wit, the equivalence of

$$\forall \mathscr{D} :\subseteq \mathscr{G}_{\underline{P}} + \mathbb{R}_{>0}; \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; \inf\{\textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\} \geq 0 \vee \sup\{\textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\} > 0$$

and

$$\forall \mathscr{D} :\subseteq \mathscr{G}_{\underline{P}}; \quad \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{D}}; \quad \sup\{\textstyle\sum_{g:\mathscr{D}} \lambda_g \cdot g\} \geq 0.$$

The last of these expressions is also encountered in the derivation of the criterion for avoiding sure loss. This means that for (sets of desirable gambles corresponding to) lower previsions, the concepts of avoiding partial loss coincides with the concept of avoiding sure loss. This is not surprising, as the difference between both concepts lies in borderline behavior, which previsions ignore (cf. $(1.18)_{37}$ and below). So for previsions this unique concept of avoiding sure loss is formalized by introducing a predicate $\mathrm{asl} : \mathscr{P}\mathscr{K} \to \mathbb{B}$:

$$\begin{aligned} \mathrm{asl}\,\underline{P} \Leftrightarrow \forall \mathscr{N} :\subseteq \mathscr{K}; \\ \forall \lambda : (\mathbb{R}_{>0})^{\mathscr{N}}; \\ \sup\{\textstyle\sum_{g:\mathscr{N}} \lambda_g \cdot G_{\underline{P}} g\} \geq 0. \end{aligned} \tag{1.25}$$

The set of lower previsions that avoid sure loss is then written $(\mathscr{P}\mathscr{K})_{\mathrm{asl}}$.

As the natural extension of a set of desirable gambles that incurs a sure loss is the set $\mathscr{L}_\Omega$, the condition above is necessary for letting the natural extension of $\underline{P}$ differ from $+\infty$, an unreasonable buying price in any case for any gamble.

Consider a lower prevision $\underline{P}$ on $\mathscr{K}$ that avoids sure loss; important consequences of this property are: (let $f : \mathscr{K}$)

Walley [1991, §2.4.7$_{71}$] mentions a large number of consequences of avoiding sure loss.

$$\text{Avoiding sure loss:} \qquad \underline{P} f \leq \sup\{f\}, \tag{1.26}$$

$$\text{Upper dominates lower:} \quad -f \in \mathscr{K} \;\Rightarrow\; \underline{P} f \leq -\underline{P}(-f) = \overline{P} f. \tag{1.27}$$

These are derived from (1.25) with specific choices for $\mathscr{N}$ and λ.

To finish this discussion, let us return to this section's introduction: My lower probability for a white Christmas must – if I want to avoid sure loss – be smaller than my upper probability, which in turn must not exceed 1. So I now choose them to be -1 and 1, respectively. These are still patently inconsistent choices and must be corrected. In this

subsection we have derived the criterion for lower previsions that, if satisfied, ensures that an uncertainty model can be corrected (made consistent) or extended to a model that specifies prices that are nowhere infinite. In the next section, we take a better look at what this consistency is and how it can be guaranteed.

1.2.5 *Coherence*

The main reference for this subsection is Walley [1991, §2.5$_{72}$].

A lower prevision that is invariant under natural extension, i.e., does not have to be corrected, satisfies a form of internal consistency. Formally, a lower prevision $\underline{P}$ on $\mathscr{P}\mathscr{K}$ is invariant under natural extension whenever it holds for all f in $\mathscr{K}$ that $\underline{P}f = \mathrm{lce}_{\underline{P}}f$. This criterion can be reformulated; we respectively start by writing its double negation and recalling that only upwards corrections are possible:

$$\neg\big(\underline{P}f \neq \mathrm{lce}_{\underline{P}}f\big) \Leftrightarrow \neg\big(\underline{P}f < \mathrm{lce}_{\underline{P}}f\big)$$

we next apply the definition of $\mathrm{lce}_{\underline{P}}$ (see somewhat above (1.23)$_{39}$) and reformulate the result:

$$\begin{aligned}&\Leftrightarrow \neg\big(\underline{P}f < \sup\{\alpha:\mathbb{R} \mid \exists g:\mathscr{R}_{\underline{P}}\,;\, f-\alpha \geq g\}\big)\\ &\Leftrightarrow \neg\big(\exists \alpha:\mathbb{R}\,;\, \underline{P}f < \alpha \wedge (\exists g:\mathscr{R}_{\underline{P}}\,;\, f-\alpha \geq g)\big)\end{aligned}$$

we continue by grouping the quantifiers, pulling the negation through them, and subsequently spotting the (material) implication:

$$\begin{aligned}&\Leftrightarrow \forall g:\mathscr{R}_{\underline{P}}\,;\forall \alpha:\mathbb{R}\,;\, \underline{P}f \geq \alpha \vee \neg(f-\alpha \geq g)\\ &\Leftrightarrow \forall g:\mathscr{R}_{\underline{P}}\,;\forall \alpha:\mathbb{R}\,;\, f-\alpha \geq g \Rightarrow \underline{P}f \geq \alpha\end{aligned}$$

now we replace the pointwise inequality by a plain one using an infimum; after some more rewriting we finish by spotting the marginal gamble:

$$\begin{aligned}&\Leftrightarrow \forall g:\mathscr{R}_{\underline{P}}\,;\forall \alpha:\mathbb{R}\,;\, \alpha \leq \inf\{f-g\} \Rightarrow \alpha \leq \underline{P}f\\ &\Leftrightarrow \forall g:\mathscr{R}_{\underline{P}}\,;\inf\{f-g\} \leq \underline{P}f\\ &\Leftrightarrow \forall g:\mathscr{R}_{\underline{P}}\,;\sup\{g-G_{\underline{P}}f\} \geq 0.\end{aligned}$$

Whenever a lower prevision both avoids sure loss and satisfies the criterion above, it is called coherent; the rationality criterion corresponding to this case is formalized using the predicate $\mathrm{coh}:\mathscr{P}\mathscr{K} \to \mathbb{B}$ with

$$\begin{aligned}\mathrm{coh}\,\underline{P} \Leftrightarrow \mathrm{asl}\,\underline{P} \wedge \forall f:\mathscr{K}\,;&\\ &\forall \mathscr{N} :\Subset \mathscr{K}_{\neq f}\,;\\ &\forall \lambda:(\mathbb{R}_{>0})^{\mathscr{N}}\,;\\ &\qquad \sup\Big\{\textstyle\sum_{g:\mathscr{N}} \lambda_g \cdot G_{\underline{P}}g - G_{\underline{P}}f\Big\} \geq 0,\end{aligned} \tag{1.28}$$

which is obtained after some back-expansion. Note that we can work with $\mathscr{K}_{\neq f}$ instead of $\mathscr{K}$ because $\underline{P}$ avoids sure loss. The set of coherent lower previsions is written as $(\mathscr{P}\mathscr{K})_{\text{coh}}$.

We have already assumed that it is reasonable to require any lower prevision to avoid sure loss. On top of that, we assume that it is also reasonable to require all lower previsions to be coherent, i.e., internally consistent in the sense that, for no gamble whatsoever, the supremum buying price can be corrected upwards. In fact, for the rest of this thesis, coherence is the strongest universal requirement. For specific problems, when evidence is available or reasonable assumptions can be made, criteria based on this can be used to further restrict the set of lower previsions that are considered an acceptable uncertainty model for the problem at hand.

Now consider a coherent lower prevision $\underline{P}$ on $\mathscr{K}$; three important consequences of this property are: (let $f, g : \mathscr{K}^2$ and $\lambda : \mathbb{R}_{\geq 0}$)

$$\text{Accepting sure gains:} \qquad \underline{P}f \geq \inf\{f\}, \tag{1.29}$$

$$\text{Nonnegative homogeneity:} \qquad \lambda \cdot f \in \mathscr{K} \;\Rightarrow\; \underline{P}(\lambda \cdot f) = \lambda \cdot \underline{P}f, \tag{1.30}$$

$$\text{Superadditivity:} \qquad f + g \in \mathscr{K} \;\Rightarrow\; \underline{P}(f+g) \geq \underline{P}f + \underline{P}g. \tag{1.31}$$

They can be derived from $(1.28)_{\curvearrowright}$ with appropriate choices for f, $\mathscr{N}$, and λ. Their importance is mainly due to the fact that these properties are equivalent to coherence $(1.28)_{\curvearrowright}$ on $\mathscr{K}$ that are linear spaces, such as $\mathscr{L}_\Omega$. (Many definitions simplify when $\mathscr{K}$ has a special structure.) Note that positive homogeneity and superadditivity combine into (let additionally $\mu : \mathbb{R}_{\geq 0}$)

Superlinearity:

$$\lambda \cdot f + \mu \cdot g \in \mathscr{K} \;\Rightarrow\; \underline{P}(\lambda \cdot f + \mu \cdot g) \geq \lambda \cdot \underline{P}f + \mu \cdot \underline{P}g. \tag{1.32}$$

Upper previsions that are conjugate to a coherent lower prevision have similar properties; most notably, they are sublinear: (1.32), but with $\underline{P}$ replaced by $\overline{P}$ and the inequality reversed.

Walley [1991, §2.6.1$_{76}$] mentions a large number of consequences of coherence.

Other consequences are: (now let $\mu : \mathbb{R}$)

$$\text{Monotonicity:} \qquad f \leq g \;\Rightarrow\; \underline{P}f \leq \underline{P}g, \tag{1.33}$$

$$\text{Normedness:} \qquad (\Omega\,;\mu) \in \mathscr{K} \;\Rightarrow\; \underline{P}(\Omega\,;\mu) = \mu, \tag{1.34}$$

$$\text{Constant additivity:} \qquad f + \mu \in \mathscr{K} \;\Rightarrow\; \underline{P}(f+\mu) = \underline{P}f + \mu, \tag{1.35}$$

$$\text{Mixed subadditivity:} \qquad (f+g, -g) \in \mathscr{K}^2 \;\Rightarrow\; \underline{P}(f+g) \leq \underline{P}f + \overline{P}g. \tag{1.36}$$

Note that normedness, positive homogeneity, and constant additivity allow for easy natural extension of $\mathscr{K}$ to constant gambles and gambles of the form $\lambda \cdot f + \mu$.

Using natural extension, the lower probability modeling my uncertainty about the occurrence of white Christmases specified at the end of

the previous subsection is corrected from -1 to 0; the upper probability 1 does not change. This results in the pair $0,1$ of coherent lower and upper probabilities: they do not imply any commitment whatsoever about taking or not taking a bet on this issue. By coincidence (honestly), this pair is an example of an interesting type of coherent lower previsions: the vacuous ones.

Vacuous lower previsions imply a total lack of commitment and thus model complete ignorance. Given a set of gambles $\mathscr{K}$, *the* vacuous lower prevision is defined by

$$\underline{P}^{\Omega} := f : \mathscr{K} ; \inf\{f\}. \tag{1.37}$$

More generally, we can also model ignorance relative to some nonempty subset A of the possibility space Ω:

$$\underline{P}^{A} := f : \mathscr{K} ; \inf\{f_A\}. \tag{1.38}$$

Use such a model to express that the *only* thing you know (believe) is that the event A will occur. A peculiar subset of these – the vacuous lower previsions relative to singletons, called degenerate previsions – actually model a *lack* of ignorance; let $\omega : \Omega$, then

$$P^{\omega} := f : \mathscr{K} ; f\omega; \tag{1.39}$$

it models an absolute conviction that ω will occur.

Now that we have translated all the concepts introduced for sets of desirable gambles to the context of lower previsions, we illustrate these concepts with a concrete example in the next subsection.

1.2.6 *An example: extending a moment*

Consider a problem where the possibility space is the unit interval, in other words, let $\Omega := [0,1]$. Assume that only the k-th raw moment $m_k :]0,1[$ is known, where k is a nonzero natural number. We wish to know what this tells us about the other raw moments. (So we only extend to part of all possible gambles.)

Miranda et al. [2006, 2007, 2008*b*,*a*] report on a more far-reaching example, concerning extension from all raw moments and the relationship with distribution functions.

The information we have about the problem can be modeled by

(i) a set of gambles $\mathscr{K} := \{v_k, -v_k\}$,
where $v_\alpha := \omega : \Omega ; \omega^\alpha$ for any nonnegative real number α,

(ii) a so-called self-conjugate lower prevision $\underline{P}$ on $\mathscr{K}$
with $\underline{P} v_k = m_k$ and $\overline{P} v_k = -\underline{P}(-v_k) = m_k$.

The information we have about the other raw moments can then be made explicit by extending $\underline{P}$ to all gambles $\pm v_\ell$ (where $\ell : \mathbb{N}$) using the procedure of natural extension. (Readers less interested in the technical details of the derivation can jump to the discussion of the results just below $(1.49)_{46}$.)

First, we show that $\underline{P}$ avoids sure loss; for this problem, $(1.25)_{40}$ can

be rewritten as

$$\begin{aligned}\operatorname{asl}\underline{P} &\Leftrightarrow \forall \lambda,\mu:(\mathbb{R}_{\geq 0})^2;\sup\{\lambda\cdot(v_k-m_k)+\mu\cdot(-v_k+m_k)\}\geq 0\\ &\Leftrightarrow \forall \nu:\mathbb{R};\sup\{\nu\cdot(v_k-m_k)\}\geq 0,\end{aligned} \tag{1.40}$$

which holds for $\nu\geq 0$ as $v_k 1 = 1$ and for $\nu\leq 0$ as $v_k 0 = 0$. So $\underline{P}$ can be extended to a coherent lower prevision $\underline{E}$ on $\{\pm v_n \mid n:\mathbb{N}\}$.

Do we actually need to correct $\underline{P}$, i.e., is it a coherent lower prevision or not? Because $\underline{P}$ avoids sure loss, we can rewrite $(1.28)_{41}$ for this problem to

$$\begin{aligned}\operatorname{coh}\underline{P} &\Leftrightarrow \begin{cases}\forall \lambda:\mathbb{R}_{>0};\sup\{\lambda\cdot(-v_k+m_k)-(v_k-m_k)\}\geq 0\\ \forall \mu:\mathbb{R}_{>0};\sup\{\mu\cdot(v_k-m_k)-(-v_k+m_k)\}\geq 0\end{cases}\\ &\Leftrightarrow \begin{cases}\forall \lambda:\mathbb{R}_{>0};\sup\{-(\lambda+1)\cdot(v_k-m_k)\}\geq 0\\ \forall \mu:\mathbb{R}_{>0};\sup\{(\mu+1)\cdot(v_k-m_k)\}\geq 0,\end{cases}\end{aligned} \tag{1.41}$$

which holds for the λ-case as $v_k 0 = 0$ and for the μ-case as $v_k 1 = 1$. So $\underline{P}$ is coherent, which means that $\underline{E}v_k := \underline{P}v_k = m_k$ and $\overline{E}v_k := \overline{P}v_k = m_k$. Actually, the condition (1.41) for coherence is equivalent to (1.40), the one for avoiding sure loss; this is typical for self-conjugate previsions (cf. §1.2.8$_{48}$ on so-called linear previsions).

There are two other gambles for which calculating the natural extension is trivial: When considering the case $\ell = 0$, we see that v_0 is the constant function I^{Ω}. Therefore, as $\underline{E}$ is coherent and therefore normed, this forces $\underline{E}v_0 := 1$ and $\overline{E}v_0 := 1$.

To obtain the natural extension for the other gambles of interest, we are going to use a formula that is equivalent to the formula $(1.23)_{39}$ we derived in §1.2.3$_{38}$ [Walley 1991, §3.1.3$_{124}$]: (we use the same notation as in $(1.23)_{39}$)

$$\operatorname{lce}_{\underline{P}} f = \sup\left\{\inf\left\{f-\textstyle\sum_{g:\mathscr{N}}\lambda_g\cdot G_{\underline{P}}g\right\}\middle|\,\mathscr{N}:\Subset\mathscr{K};\lambda:(\mathbb{R}_{>0})^{\mathscr{N}}\right\}.$$

Taking the specific set of gambles $\mathscr{K}$ we currently consider and the specific form of the gambles to which we wish to extend into account, the lower and upper prevision of gambles v_ℓ are then defined by

$$\underline{E}v_\ell := \operatorname{lce}_{\underline{P}} v_\ell = \sup_{\mu:\mathbb{R}}\big(\mu\cdot m_k+\min\{v_\ell-\mu\cdot v_k\}\big), \tag{1.42}$$

$$\overline{E}v_\ell := -\operatorname{lce}_{\underline{P}}(-v_\ell) = \inf_{\mu:\mathbb{R}}\big(\max\{v_\ell+\mu\cdot v_k\}-\mu\cdot m_k\big). \tag{1.43}$$

For $\operatorname{lce}_{\underline{P}} v_\ell$, the right-hand side is 0 when only considering nonpositive μ, for $-\operatorname{lce}_{\underline{P}}(-v_\ell)$, the right-hand side is 1 when only considering nonnegative μ. So we can rewrite (1.42) and (1.43) as

$$\underline{E}v_\ell := \max\{0,\sup_{\lambda:\mathbb{R}_{>0}}(\lambda\cdot m_k+\min\{v_\ell-\lambda\cdot v_k\})\}, \tag{1.44}$$

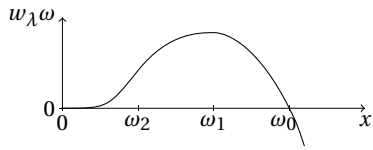
$$\overline{E}v_\ell := \min\{1,\inf_{\lambda:\mathbb{R}_{>0}}(\lambda\cdot m_k+\max\{v_\ell-\lambda\cdot v_k\})\}. \tag{1.45}$$

A sign analysis of the function $w_\lambda := v_\ell - \lambda \cdot v_k$ and its derivatives $\mathrm{D}w_\lambda$ and $\mathrm{D}^2 w_\lambda$ allows us to characterize the functions $\lambda : \mathbb{R}_{>0} ; \min\{w_\lambda\}$ and $\lambda : \mathbb{R}_{>0} ; \max\{w_\lambda\}$. This in turn allows us to simplify (1.44) and (1.45). We now treat the cases for which ℓ is respectively strictly smaller or larger than k.

Illustrating derivation: $\mathrm{D}\sin = \cos$, $\mathrm{D}^2(x : \mathbb{R} ; x^2) = 2$.

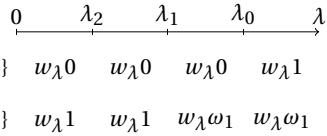
For $\ell < k$, the somewhat involved analysis is summarized by a sketch of the function w_λ and a table for its minimum and maximum. Important quantities in this analysis are the nontrivial zeroes of w_λ and its derivatives:

$$\begin{aligned} \omega_0 &:= \left(\tfrac{1}{\lambda}\right)^{\frac{1}{k-\ell}}, & w_\lambda \omega_0 &= 0, \\ \omega_1 &:= \left(\tfrac{1}{\lambda} \cdot \tfrac{\ell}{k}\right)^{\frac{1}{k-\ell}}, & (\mathrm{D}w_\lambda)\omega_1 &= 0, \\ \omega_2 &:= \left(\tfrac{1}{\lambda} \cdot \tfrac{\ell \cdot (\ell-1)}{k \cdot (k-1)}\right)^{\frac{1}{k-\ell}}, & (\mathrm{D}^2 w_\lambda)\omega_2 &= 0. \end{aligned}$$



Also important are the threshold values of λ for which such a zero is equal to 1:

$$\begin{aligned} \omega_0 = 1 &\Rightarrow \lambda = \lambda_0 := 1, \\ \omega_1 = 1 &\Rightarrow \lambda = \lambda_1 := \tfrac{\ell}{k}, \\ \omega_2 = 1 &\Rightarrow \lambda = \lambda_2 := \tfrac{\ell \cdot (\ell-1)}{k \cdot (k-1)}. \end{aligned}$$



This analysis allows us to get to conclusions from (1.44) and (1.45) for the case $\ell < k$:

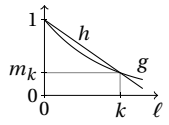
$$\begin{aligned} \underline{E} v_\ell &:= \max\{0, \sup_{\lambda :]0,\lambda_0]} \lambda \cdot m_k, \sup_{\lambda : \mathbb{R}_{>\lambda_0}} (\lambda \cdot m_k + 1 - \lambda)\} \\ &= \max\{0, m_k, m_k\} \qquad (1.46) \\ &= m_k, \\ \bar{E} v_\ell &:= \min\{1, \inf_{\lambda :]0,\lambda_1]} (\lambda \cdot m_k + 1 - \lambda), \inf_{\lambda : \mathbb{R}_{>\lambda_1}} (\lambda \cdot m_k + w_\lambda \omega_1)\} \\ &= \min\{1, 1 - \tfrac{\ell}{k}(1 - m_k), (m_k)^{\ell/k}\} \qquad (1.47) \\ &= (m_k)^{\ell/k}. \end{aligned}$$

In (1.47), two steps cannot be verified on sight:

(i) The equality of $\inf_{\lambda : \mathbb{R}_{>\lambda_1}} (\lambda \cdot m_k + w_\lambda \omega_1)$ and $(m_k)^{\ell/k}$. This infimum, which is attained in $\lambda := \lambda_1 \cdot (m_k)^{(l-k)/k}$, is found with a sign analysis of $\lambda \cdot m_k + w_\lambda \omega_1$ and its derivatives, where

$$w_\lambda \omega_1 = \left(\tfrac{1}{\lambda} \cdot \tfrac{\ell}{k}\right)^{\frac{k}{k-\ell}} - \lambda \cdot \left(\tfrac{1}{\lambda} \cdot \tfrac{\ell}{k}\right)^{\frac{\ell}{k-\ell}}.$$

(ii) That the function $h := \ell : \mathbb{N}_{\leq k} ; 1 - \frac{\ell}{k} \cdot (1 - m_k)$ dominates the function $g := \ell : \mathbb{N}_{\leq k} ; (m_k)^{\ell/k}$. This can be proven with a sign analysis of their difference; it is made acceptable with a plot.



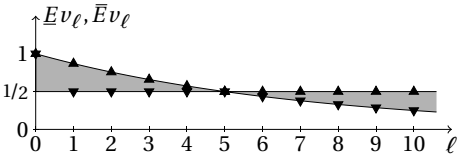
For $\ell > k$, similar results hold as the one in the previous case because of symmetry considerations. These results can be obtained from the above by replacing the sketch by its negative, switching minimum and

maximum in the table, and – in the expressions for the zeroes – switching k and ℓ and replacing λ by $1/\lambda$. Amongst other things, this results in a minimum at $\varpi_1 := (\lambda \cdot \frac{k}{\ell})^{1/\ell-k}$. So this gets us the following conclusions from $(1.44)_{44}$ and $(1.45)_{44}$ for the case $\ell > k$:

$$\begin{aligned}\underline{E}v_\ell &:= \max\{0, \sup_{\lambda:]0,\lambda_1]}(\lambda \cdot m_k + 1 - \lambda), \sup_{\lambda:\mathbb{R}_{>\lambda_1}}(\lambda \cdot m_k + w_\lambda \varpi_1)\} \\ &= \max\{0, 1 - \tfrac{\ell}{k}(1 - m_k), (m_k)^{\ell/k}\} \\ &= (m_k)^{\ell/k}, \end{aligned} \tag{1.48}$$

$$\begin{aligned}\overline{E}v_\ell &:= \min\{1, \inf_{\lambda:]0,\lambda_0]}\lambda \cdot m_k, \inf_{\lambda:\mathbb{R}_{>\lambda_0}}(\lambda \cdot m_k + 1 - \lambda)\} \\ &= \min\{1, m_k, m_k\} \\ &= m_k. \end{aligned} \tag{1.49}$$

The results in (1.46)–$(1.49)_{45–46}$ can be nicely summarized with a plot (we chose k to be 5 and m_5 to be $1/2$). The values $\underline{E}v_\ell$ are indicated using a downwards-pointing triangle ▾, and the values $\overline{E}v_\ell$ by an upwards-pointing triangle ▴. Notice that $\underline{E}v_0 = \overline{E}v_0 = 1$.



Nothing in our analysis depended on ℓ (or k) being integer-valued. The same results would be obtained for $\ell : \mathbb{R}_{\geq 0}$, which is why the graph also has curves for the corresponding values $\underline{E}v_\ell$ and $\overline{E}v_\ell$.

Imprecision as a measure of (lack of) information is not related to the measure of information used in information theory, which is a difference of two Shannon entropies [1948].

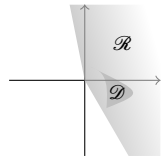
The difference between the lower and upper prevision of a gamble is often called the imprecision. Lower values for lower previsions and higher values for upper previsions indicate less commitment. This commitment reflects the information available about the gamble under scrutiny (or at least it should). The imprecision of a gamble can therefore be seen as a measure for (lack of) the information available about that gamble. In the plot, we draw attention to the fact that the imprecision increases as ℓ moves away from $k := 5$, by shading the area between the plots of the functions $\ell : \mathbb{R}_{\geq 0}\,; \underline{E}v_\ell$ and $\ell : \mathbb{R}_{\geq 0}\,; \overline{E}v_\ell$.

In the next subsection, we show that there are other possible extensions than the natural extension used here. As was the case here, the concept of commitment will still be used to give an interpretation to these extensions.

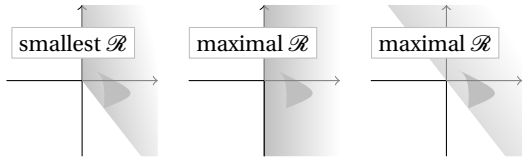
1.2.7 *Least and maximally committal extensions*

Let us return for a while to the land of sets of desirable gambles we first visited in §1.2.1$_{34}$. We saw that whenever some set of desirable gambles $\mathscr{D} :\subseteq \mathscr{L}_\Omega$ avoided partial loss, its natural extension – the coherent set of desirable gambles $\mathscr{R}_\mathscr{D}$ (cf. $(1.10)_{35}$) – took the form of a convex cone containing, apart from the origin, the entire nonnegative orthant $(\mathscr{L}_\Omega)_{\geq 0}$, but no gamble from the nonpositive orthant $(\mathscr{L}_\Omega)_{\leq 0}$. Any

other convex cone $\mathscr{R} :\subset \mathscr{L}_\Omega$ with these properties and containing $\mathscr{D}$ is a coherent extension of $\mathscr{D}$.



The many cones that are coherent extensions of $\mathscr{D}$ can be partially ordered; we can identify both a unique smallest cone (contained in all coherent extensions) and a set of maximal cones (not contained in any other extension). These correspond to the so-called least committal extension and maximally committal extensions, respectively. An illustration is given on the side.



The unique smallest cone corresponds to the already encountered natural extension $(1.10)_{35}$. We also already encountered the natural extension of a lower prevision $\underline{P}$ on $\mathscr{K}$ in §1.2.3$_{38}$, where we introduced the least committal extension function $\text{lce}_{\underline{P}}$, whose name is derived from an alternative name for natural extension: least committal extension. This is an apt name, as the natural extension of $\underline{P}$ assigns the lowest supremum acceptable buying price one is implicitly committed to when specifying $\underline{P}$.

It is interesting to note that the least committal extension procedure is transitive [this is implicit in Walley 1991, §3.4.1$_{136}$]: Take any two gambles f and g on Ω and define the lower prevision $\underline{E}$ on $\mathscr{K} \cup \{f\}$ by $\underline{E}_{\mathscr{K}_{\neq f}} = \underline{P}$ and $\underline{E}f = \text{lce}_{\underline{P}} f$, then $\text{lce}_{\underline{P}} g = \text{lce}_{\underline{E}} g$. This also works, of course, after extension to multiple gambles.

The maximal cones correspond to the maximally committal extension procedure, which is named as it is, because it assigns the highest supremum acceptable buying price that does not result in a sure loss. For any lower prevision $\underline{P}$ on $\mathscr{K}$ and any gamble f in $\mathscr{K}$, this price is calculated using the most committal extension function, defined as the least upper bound of the prices leading to a partial and thus a sure loss (cf. §1.2.4$_{39}$); we start from $(1.24)_{40}$ and add a tentative marginal gamble for f to the gamble of $\mathscr{R}_{\underline{P}}$ under scrutiny:

$$\text{mce}_{\underline{P}} f := \inf\{\beta : \mathbb{R} \mid \exists g : \mathscr{R}_{\underline{P}} ; \inf\{g + (f - \beta)\} < 0 \wedge \sup\{g + (f - \beta)\} \leq 0\}$$

we rewrite the inequalities involving the supremum and infimum by using pointwise inequalities and do some back-expansion:

$$= \inf\left\{\beta : \mathbb{R} \;\middle|\; \begin{array}{l} \exists \mathscr{N} :\Subset \mathscr{K} ; \\ \exists \lambda : (\mathbb{R}_{>0})^{\mathscr{N}} ; \\ \qquad \beta - f \geq \sum_{g:\mathscr{N}} \lambda_g \cdot G_{\underline{P}} g \end{array}\right\}. \tag{1.50}$$

It is important to note that the maximally committal extension procedure is *not* transitive. It assigns the highest reasonable supremum

buying price consistent with the judgements $\underline{P}$ *and* a specific sequence of gambles to be extended; different sequences can result in different extended lower previsions. The correct way to apply this procedure is in a stepwise fashion: Take any two gambles f and g on Ω and in a first step define $\underline{E}:\mathscr{K}\cup\iota f$ by $\underline{E}_{\mathscr{K}_{\neq f}}=\underline{P}$ and $\underline{E}f=\mathrm{mce}_{\underline{P}}f$, then in the next step the extension to g *must* be defined as $\mathrm{mce}_{\underline{E}}g$.

It follows from $(1.23)_{39}$ and $(1.50)_{\cap}$ that $\mathrm{mce}_{\underline{P}}f=-\mathrm{lce}_{\underline{P}}(-f)$ for any gamble f on Ω. This implies that $\mathrm{mce}_{\underline{P}}$ is also the function that returns the natural extension for upper previsions [Walley 1991, §3.1.3$_{124}$]; to wit, let $\underline{E}:\mathscr{P}\mathscr{L}_\Omega$ be the natural extension of $\underline{P}$, then, by conjugacy $(1.14)_{36}$,

$$\overline{E}f=-\underline{E}(-f)=-\mathrm{lce}_{\underline{P}}(-f)=\mathrm{mce}_{\underline{P}}f.$$

As will be justified in §1.2.9$_{50}$, maximally committal extension can also be called extremal extension and any possible extended lower prevision an extremal extension.

Using extremal extension and starting from our choices in the last paragraph of §1.2.4$_{39}$ – the pair $-1,1$ of lower and upper probabilities for white Christmases –, we can obtain $1,1$ or $0,0$ as pairs of lower and upper probabilities for my uncertainty about the occurrence of white Christmases. The former pair implies that I should always bet on a white Christmas happening, the latter that I should always bet against this.

So, we now know there are other extension procedures than natural or least committal extension. We also know some more details about one of these procedures, the maximally committal or extremal extension. The question is: why did we bother to introduce the extremal extension? The next subsection gives an answer.

1.2.8 *Linear previsions*

The main reference for this subsection is Walley [1991, §2.8$_{86}$].

In §1.2.5$_{41}$, we discovered the coherent lower previsions as the ones that avoided sure loss and were invariant under natural or least committal extension. Being naturally curious, we like to find out what the properties of the lower previsions are that are additionally invariant under maximally committal extension.

In the previous subsection, we saw that the function used for calculating the maximally committal extension is actually also the natural extension function for upper previsions. So a coherent lower prevision that is invariant under maximally committal extension is at the same time a coherent upper prevision. It then follows from $(1.14)_{36}$ that this prevision is self-conjugate (for gambles whose negative is also in the prevision's domain). Moreover, this prevision is both superlinear and sublinear (cf. $(1.32)_{42}$ and below), so it is linear, a property that will also give it its name.

We cast a very general definition for linear lower previsions from the

literature [Walley 1991, §2.8.1$_{86}$] in the predicate $\text{lin}:\mathscr{P}\mathscr{K} \to \mathbb{B}$:

$$\begin{aligned}\text{lin}\,\underline{P} \Leftrightarrow \forall \mathcal{N} :\Subset \mathscr{K};\\ \forall \lambda : \mathbb{R}^{\mathcal{N}};\\ \sup\{\textstyle\sum_{g:\mathcal{N}} \lambda_g \cdot G_{\underline{P}} g\} \geq 0.\end{aligned} \tag{1.51}$$

Miranda [2008*a*] nicely illustrates the subtleties involved in the definition of linear previsions.

Because the coefficients are allowed to take on any value, this criterion is more restrictive than coherence (1.28)$_{41}$ and avoiding sure loss (1.25)$_{40}$. The set of linear lower previsions is then written as $(\mathscr{P}\mathscr{K})_{\text{lin}}$.

Consider a linear lower prevision P on $\mathscr{K}$; as alluded to in the beginning of this subsection, the most important consequences of linearity – that are not a consequence of coherence – are (let $f, g : \mathscr{K}^2$)

$$\text{Self-conjugacy:} \quad -f \in \mathscr{K} \;\Rightarrow\; Pf = -P(-f), \tag{1.52}$$

$$\text{Additivity:} \quad f+g \in \mathscr{K} \;\Rightarrow\; P(f+g) = Pf + Pg. \tag{1.53}$$

Note that due to the extra properties of self-conjugacy and additivity, extension to the linear span of $\mathscr{K}$ is easy. For example, if the possibility space Ω is finite and $\Omega \subset (\text{span}\,\mathscr{K})^\star$, extension to $\mathscr{L}_\Omega$ is immediate: a linear lower prevision is then fully specified once it is specified on the elementary events (cf. (1.2)$_{31}$ for the meaning of '$\star$').

Because these lower previsions coincide with their conjugate upper prevision, we can from now on drop 'lower' and just talk about linear previsions. They can be interpreted as giving a fair price for buying *and* selling [de Finetti 1974–1975]. Also, as in the previous paragraph, previsions known to be linear are denoted without an underbar.

In §1.2.7$_{46}$, we obtained the pairs 0, 0 and 1, 1 of lower and upper probabilities for my uncertainty about the occurrence of white Christmases as the two possible results of the extremal correction (and extension) procedure. By coincidence, both pairs are degenerate previsions, which themselves are also always linear – epitomized by the prescient lack of an underbar in their definition (1.39)$_{43}$.

Having landed at the end of this subsection, one could wonder why so much attention is being paid to linear previsions, as linearity is not a rationality criterion. Linear previsions, with their attractive mathematical properties, are – possibly in a different guise –, in a large part of the literature, the *only* uncertainty models considered. This thesis positions itself in another part of the literature, which tries to avoid implicit assumptions (that support the central role of linear previsions) and where a different balance is struck between mathematical complexity and modeling power. However, linear previsions still have a role to play, thanks to their mathematical properties and usefulness as part of an alternative representation of the information contained in a coherent lower prevision. This is the subject of the next subsection.

1.2.9 *Credal sets*

Walley [1991, §D$_{608}$] and Maaß [2003*b*, §2.2] discuss the topological structure of sets involved in modeling uncertainty more thoroughly.

Let us start this subsection with a technical intermezzo: Consider a set of gambles $\mathscr{K} :\subseteq \mathscr{L}_\Omega$. We always choose the supremum-norm topology on $\mathscr{K}$ and the topology of pointwise convergence on $\mathscr{P}\mathscr{K}$. (For finite Ω, both are equivalent to the Euclidean or metric topologies.) These choices, and (the ideas behind) the convexity and convergence theorems of Walley [1991, §2.6.4–5$_{79}$], when applied to the definitions of avoiding sure loss (1.25)$_{40}$, coherence (1.28)$_{41}$, and linearity (1.51)$_{\frown}$ show that $(\mathscr{P}\mathscr{K})_{\text{asl}}$, $(\mathscr{P}\mathscr{K})_{\text{coh}}$, and $(\mathscr{P}\mathscr{K})_{\text{lin}}$ are closed convex sets.

Now, starting from a lower prevision $\underline{P}$ on $\mathscr{K}$ that avoids sure loss, the ideas mentioned above show that the set of all linear previsions that dominate $\underline{P}$ is closed and convex, which can even be shown to be compact [Walley 1991, §3.6.1$_{145}$]. It is called the credal set of $\underline{P}$. Its definition directly leads to the introduction of the function $\mathscr{M} : (\mathscr{P}\mathscr{K})_{\text{asl}} \to \wp(\mathscr{P}\mathscr{L}_\Omega)_{\text{lin}}$ that generates credal sets:

Levi [1974] introduced credal sets (using the term 'credal state').

$$\mathscr{M}\underline{P} = \left\{P : (\mathscr{P}\mathscr{L}_\Omega)_{\text{lin}} \mid P_{\mathscr{K}} \geq \underline{P}\right\}. \tag{1.54}$$

Let f be any gamble in $\mathscr{K}$, then Walley [1991, §3.4.1$_{136}$ and §3.6.2$_{146}$] provides us with a sort of converse of the equation above under the form of a lower envelope theorem:

$$\begin{aligned} \text{lce}_{\underline{P}}\, f &= \min\nolimits_{P:\mathscr{M}\underline{P}} Pf \\ &= \min\nolimits_{P:\text{ext}(\mathscr{M}\underline{P})} Pf, \end{aligned} \tag{1.55}$$

Illustrating the extractor of extreme points: $\text{ext}[0,1] = \mathbb{B}$.

where ext is the function that extracts, from a compact convex set, the set of extreme points, of which convex mixtures (and limits thereof) can be used to define every element of the compact convex set. For *coherent* previsions $\underline{P}$ on $\mathscr{K}$ we get a real converse to (1.54), as then $\text{lce}_{\underline{P}}\, f = \underline{P} f$ for all f [Walley 1991, §3.3.3$_{134}$ and §3.4.1$_{136}$]. This one-to-one relationship between (the extreme points of) credal sets and coherent lower previsions means that they can be used as equivalent uncertainty models.

The extreme points of the credal set are, at least for finite possibility spaces, the previsions resulting from extremal extension. This is so by construction: To calculate an extremal extension, apply the maximally committal extension procedure (cf. §1.2.7$_{46}$) to all the indicators of elementary events in some well-defined sequence and then use linearity to further calculate the extension to all gambles. Now, let $E : (\mathscr{P}\mathscr{L}_\Omega)_{\text{lin}}$ be an extremal extension of $\underline{P}$ and let ω and ϖ be the first two of its defining sequence of elementary events in Ω. Then $E\omega = \max_{P:\mathscr{M}\underline{P}} P\omega$; if E is not the unique such extension with this property, then we can use $E\varpi = \max_{P:\mathscr{M}\underline{P} \wedge P\omega = E\omega} P\varpi$, and so forth.

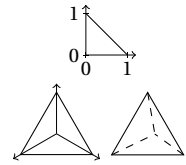
Working with coherent lower previsions through their credal set is

not only mathematically useful: For finite possibility spaces Ω, graphical illustrations can be based on this correspondence. We have already seen that linear previsions defined on a domain that includes all elementary events are then fully defined by the values they take in these elementary events. So assume Ω is finite; any linear prevision P on $\mathscr{L}_\Omega$ can then be represented by P_Ω, a point of $\mathbb{R}^{|\Omega|}$. However, not all points of $\mathbb{R}^{|\Omega|}$ represent linear previsions; the set of all linear previsions satisfies the following additional restrictions:

(i) avoiding sure loss $(1.26)_{40}$ implies $P_\Omega \leq 1$,
(ii) accepting sure gains $(1.29)_{42}$ implies $P_\Omega \geq 0$,
(iii) additivity $(1.53)_{49}$ and normedness $(1.34)_{42}$ imply $\sum P_\Omega = P(I^\Omega) = 1$.

So for any finite possibility space Ω, the set of all linear previsions is represented by the so-called $(|\Omega|-1)$-dimensional unit simplex

$$\Delta_\Omega := \{q : \Omega \to [0,1] \mid \textstyle\sum q = 1\}. \tag{1.56}$$

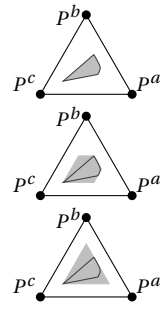


For $|\Omega| = 2$ it is a line segment, for $|\Omega| = 3$ a triangle (including its interior), and for $|\Omega| = 4$ a tetrahedron (a three-dimensional figure in a four-dimensional space whose axes we cannot visualize). Note that the unit simplex is also useful to define the vector of coefficients for a convex mixture.

Like the set of linear previsions that it represents, the unit simplex is a bounded closed convex set. Its extreme points represent the degenerate previsions. As these are the extreme points of the set of all linear previsions for finite possibility spaces [Walley 1991, §3.6.7_{149}], we can identify the unit simplex (and its elements) with the set of linear previsions (and its elements) for all intents and purposes.

We have already seen in §1.1.7_{33} that the higher a lower prevision of a gamble, the bigger the implied commitment. Whenever the uncertainty is measured reasonably, the commitments concerning a gamble should reflect the amount of information available about that gamble. So, a coherent lower prevision with higher values is a model that reflects more information. There are less linear previsions that dominate a coherent lower prevision with higher values, so smaller credal sets reflect more information.

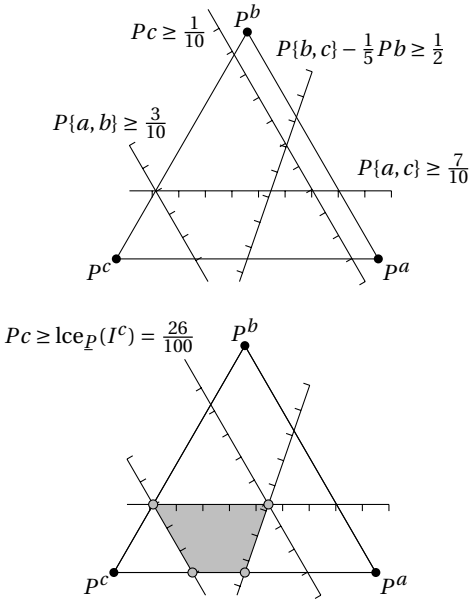
We now use the unit simplex and credal sets to illustrate that a coherent lower prevision generally contains more information than its restriction to events or elementary events. For this, let $\Omega := \{a, b, c\}$, let $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ be such that $\wp\Omega \subset \mathscr{K}^\star$, and let a coherent lower prevision $\underline{P}$ on $\mathscr{K}$ be defined by its credal set $\mathscr{M}\underline{P}$, drawn (in gray) in the top unit simplex. In the unit simplices below, we find the credal sets $\mathscr{M}\underline{P}_{\wp\Omega}$ and $\mathscr{M}\underline{P}_\Omega$ of its restriction to events and elementary events, respectively. As each of the latter two credal sets encompasses the previous one, $\underline{P}$ reflects more information than $\underline{P}_{\wp\Omega}$, which in turn reflects more information than $\underline{P}_\Omega$.



To finish this subsection, we use the unit simplex to illustrate the

extension procedures described in §1.2.3$_{38}$ and §1.2.7$_{46}$. Again, as $|\Omega| = 3$ is the case best suited for graphical illustrations, let $\Omega := \{a, b, c\}$. Furthermore, consider some set of gambles $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ and some lower prevision $\underline{P} : (\mathscr{P}\mathscr{K})_{\text{asl}}$, defined concurrently by

$$\begin{aligned} \underline{P}\{a, b\} &= \tfrac{3}{10}, \\ \underline{P}c &= \tfrac{1}{10}, \\ \underline{P}\{a, c\} &= \tfrac{7}{10}, \\ \underline{P}(I^{\{b,c\}} - \tfrac{1}{5} \cdot I^b) &= \tfrac{1}{2}. \end{aligned}$$



Each of these assignments constrains all of the possible linear extensions $P : \mathscr{M}\underline{P}$ of $\underline{P}$. These four constraints are drawn on the first simplex shown; the hairs of the lines representing the constraints point into the direction of the P that are allowed by that constraint. On a second simplex, we have drawn the constraints that result from applying natural extension; only one constraint needed correcting, $\underline{P}c = 1/10$. The credal set $\mathscr{M}\underline{P}$ that stands in a one-to-one relation with the resulting coherent lower prevision is again drawn in gray. Its extreme points $\text{ext}(\mathscr{M}\underline{P})$, the linear previsions resulting from extremal extension, are indicated using a little circle ∘.

1.3 RESTRICTING, TRANSFORMING & COMBINING UNCERTAINTY MODELS

This section is again based mainly on Walley's [1991] work, and again through him, on Williams's [1975].

In the previous sections, we have only talked about uncertainty models consisting of a single lower prevision defined on some gambles on a given possibility space. In this section, we are going to add lower previsions defined on some gambles on (a possibly transformed) *part* of this possibility space to the mix and look at the relationships between both types of models imposed by rationality.

1.3.1 *Marginal and induced previsions*

The first kind of restricted uncertainty models we look at are the so-called marginal previsions (the term 'marginal' here is in no way related to marginal or marginally desirable gambles). It is applicable when the possibility space Ω can be written as a cartesian product $\mathscr{X} \times \mathscr{Y}$ of sets $\mathscr{X}$ and $\mathscr{Y}$, which are possibility spaces in their own right.

A gamble $\tilde{f}$ on $\mathscr{X}\times\mathscr{Y}$ is called the cylindrical extension of a gamble f on $\mathscr{X}$ if $\tilde{f}=f$ (or, by explicitly writing out the pointwise extension, if $\forall y:\mathscr{Y}\,;\tilde{f}(\bullet,y)=f$). Given some set of gambles $\mathscr{K}:\subseteq\mathscr{L}_\Omega$, the set of gambles in $\mathscr{L}_\mathscr{X}$ that have a cylindrical extension in $\mathscr{K}$ is

$$\mathscr{K}_\mathscr{X}:=\{f:\mathscr{L}_\mathscr{X}\mid\exists\tilde{f}:\mathscr{K}\,;\tilde{f}=f\}. \tag{1.57}$$

Illustration with $\mathscr{X}:=\{a,b\}$ and $\mathscr{Y}:=\{c,d\}$:

$$f:=\begin{pmatrix}f\,a\\f\,b\end{pmatrix},$$

$$\tilde{f}:=\begin{pmatrix}\tilde{f}(a,c)&\tilde{f}(a,d)\\\tilde{f}(b,c)&\tilde{f}(b,d)\end{pmatrix}=\begin{pmatrix}f\,a&f\,a\\f\,b&f\,b\end{pmatrix}.$$

Now consider a lower prevision $\underline{P}$ on $\mathscr{K}$, then its $\mathscr{X}$-marginal is the lower prevision $\underline{R}$ on $\mathscr{K}_\mathscr{X}$, which is defined for every gamble f in $\mathscr{K}_\mathscr{X}$ by $\underline{R}f=\underline{P}\tilde{f}$. The behavioral interpretation is that a desirable gamble which remains constant under the variation of some possibly influencing factor remains desirable when this factor is not considered.

When $\underline{P}$ avoids sure loss or is coherent, then this property is inherited by $\underline{R}$. This is so because

(i) if the corresponding criteria $(1.25)_{40}$ or $(1.28)_{41}$ hold for $\underline{P}$ (defined on $\mathscr{K}$), then they also hold for its restriction to the set of cylindrical extensions $\{\tilde{f}:\mathscr{K}\mid\exists f:\mathscr{K}_\mathscr{X}\,;\tilde{f}=f\}$, and thus

(ii) the criteria also hold for $\underline{R}$, as cylindrical extension is a linear, constant additive operation and also supremum-preserving, i.e., $\sup\{g\}=\sup\{\tilde{g}\}$ by construction for any gamble g on $\mathscr{X}$ and its cylindrical extension $\tilde{g}$ on $\mathscr{X}\times\mathscr{Y}$.

The idea behind marginalization can be generalized: notice that the central element above was a way to associate a gamble in $\mathscr{L}_\mathscr{Z}$ – where $\mathscr{Z}$ is some set (above, this is $\mathscr{X}$) – with a gamble in $\mathscr{L}_\Omega$ (above, this is cylindrical extension). So consider a lower prevision $\underline{P}$ on $\mathscr{K}$ and a function $\Gamma:\mathscr{L}_\mathscr{Z}\to\mathscr{L}_\Omega$, then the induced lower prevision $\underline{Q}$ on $\mathscr{K}_\mathscr{Z}$ is defined for every gamble f in $\mathscr{K}_\mathscr{Z}$ by $\underline{Q}f=\underline{P}(\Gamma f)$, where

$$\mathscr{K}_\mathscr{Z}:=\{f:\mathscr{L}_\mathscr{Z}\mid\Gamma f\in\mathscr{K}\} \tag{1.58}$$

is the set of gambles on $\mathscr{Z}$ with an associated gamble in $\mathscr{K}$. When the function Γ is linear, constant additive, and supremum-preserving (as was the case for cylindrical extension), this is sufficient for the properties of avoiding sure loss and coherence to be inherited.

The parallels between marginal previsions and induced previsions are strongest when Γ is defined by a surjective map $\varphi:\Omega\to\mathscr{Z}$, i.e., when the set $\mathscr{Z}$ is isomorphic to a partition of Ω. Then for each gamble f in $\mathscr{L}_\mathscr{Z}$ we have $\Gamma f:=f\circ\varphi$. We then also give the name marginal prevision to $\underline{Q}$ and the restriction of $\underline{P}$ to the set $\{\Gamma f\mid f:\mathscr{K}_\mathscr{Z}\}$ of gambles that are constant on the events in the induced partition.

Illustrating function composition: $\ln_{\mathbb{R}_{>0}}\circ\exp_\mathbb{R}=\mathrm{id}_\mathbb{R}$.

The following points are noteworthy about this last case:

(i) Being defined by a surjective map φ, Γ is linear, constant additive, and supremum-preserving, so avoiding sure loss and coherence are automatically preserved.

(ii) When φ is the projection from $\mathscr{X}\times\mathscr{Y}$ to $\mathscr{X}$, Γ becomes cylindrical extension and we find back the case that opened this subsection.

The map $\varphi^{-1}:\wp\mathscr{Z}\to\wp\Omega$ is defined for every $A:\subseteq\mathscr{Z}$ by $\varphi^{-1}A=\{\omega:\Omega\,|\,\varphi\omega\in A\}$.

(iii) Take $A:\subseteq\mathscr{Z}$, then $\Gamma I^A=I^{\varphi^{-1}A}$, where $\varphi^{-1}A$ is the inverse image of A in Ω, then $\underline{Q}A=\underline{P}(\varphi^{-1}A)$, an expression quite often seen in the literature [see, e.g., Burrill 1972, §7-8_{136}].

Conceptually, marginal (and induced) previsions are relatively simple when compared to conditional previsions, the next type of restricted uncertainty models we are going to take a look at.

1.3.2 *Contingent, updated & conditional lower previsions*

Consider a possibility space Ω. In many situations uncertainty models relative to some nonempty conditioning event B of Ω – where B takes the role of Ω – are useful, be it because these may be easier to assess or necessary in the analysis of the problem at hand.

Our workhorse uncertainty model is the prevision, so consider a lower prevision $\underline{P}(\bullet|B)$ defined on the set of gambles $\mathscr{K}_B:\subseteq\mathscr{L}_B$. (Note the explicitly different notation.) It is called either a contingent or an updated lower prevision, depending on what it models. The former refers to *current* commitments contingent on the event B happening, the latter to *current* commitments for the situation where it has become clear – due to some observation – that B has happened (but nothing more specific is learned). Contingent previsions say nothing about buying prices contingent on B *not* happening; this is formalized by specifying a price of 0 for the zero gamble on $\Omega\setminus B$, i.e., by trivial extension (cf. §0.3.4_{25}) of marginal gambles from B to Ω, as we will see later on in this subsection. We assume that contingent and updated commitments coincide [Walley 1991, §6.1.6_{287}, the updating principle]; we use the term updated prevision generically for both from now on.

The interpretation given to an updated prevision – current commitments for the situation after the observation of the occurrence of an event – makes it natural to consider sets of updated lower previsions whose corresponding observed events form a partition $\mathscr{B}:\subseteq\wp\Omega$ of the possibility space: Any experiment that allows us to observe the occurrence of some event immediately reveals the nonoccurrence of its complement; finer partitions can be obtained by chaining such experiments or performing other, more discriminatory types of experiments.

We do not give a direct treatment of infinite partitions and its many pitfalls. When they are needed, we obtain them – in Jaynesian [2003, §B.2_{662}] fashion – using limit arguments.

A set $\{\underline{P}(\bullet|B)\,|\,B:\mathscr{B}\}$ of updated lower previsions and its associated set of gambles

$$\mathscr{K}:=\{f:\mathscr{L}_\Omega\,|\,\forall B:\mathscr{B};f_B\in\mathscr{K}_B\}\tag{1.59}$$

allow us to introduce the conditional lower prevision $\underline{P}(\bullet|\mathscr{B}):\mathscr{K}^\star\to\mathscr{L}_\Omega$, which is defined for every gamble f in $\mathscr{K}$ by

$$\underline{P}(f|\mathscr{B})=\textstyle\sum_{B:\mathscr{B}}\underline{P}(f|B)\cdot I^B,\tag{1.60}$$

where we used the notational convention $\underline{P}(f|B) = \underline{P}(f_B|B)$. Note that the definition of a conditional lower prevision preserves the conjugacy of the defining (updated) lower previsions, so we can also talk about the conjugate conditional upper prevision $\overline{P}(\bullet|\mathscr{B})$ defined on $-\mathscr{K}$. Also note that an evaluated conditional lower prevision such as $\underline{P}(f|\mathscr{B})$ is a gamble; we say that it is generated from f by $\underline{P}(\bullet|\mathscr{B})$.

The conditional lower prevision $\underline{P}(\bullet|\mathscr{B})$ generates gambles that are constant on the elements of the partition $\mathscr{B}$. Such gambles are called $\mathscr{B}$-measurable and formalized by the predicate $\mathscr{B}\text{-msr}:\mathscr{L}_\Omega \to \mathbb{B}$, defined for every gamble f on Ω by

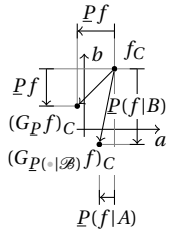
$$\begin{aligned}\mathscr{B}\text{-msr}\, f &\Leftrightarrow \forall B:\mathscr{B}; \exists \alpha_B:\mathbb{R}; f_B = \alpha_B \\ &\Leftrightarrow \forall B:\mathscr{B}; \forall(\omega,\varpi):B^2; f\omega = f\varpi.\end{aligned} \tag{1.61}$$

A so-called unconditional lower prevision $\underline{P}$ on $\mathscr{K}$, where we now take $\mathscr{K} :\subseteq \mathscr{L}_\Omega$, can also be looked at as an updated lower prevision for the trivial conditioning event Ω. The corresponding conditional lower prevision $\underline{P}(\bullet|\imath\Omega):\mathscr{K}^\star \to \mathscr{L}_\Omega$ generates constant functions, i.e., $\underline{P}(f|\imath\Omega) = \underline{P}(f|\Omega) = \underline{P}f$ for every f in $\mathscr{K}$.

Conditional lower previsions such as $\underline{P}(\bullet|\mathscr{B})$ are used as an intermediate between sets of updated lower previsions and sets of desirable gambles. Recall that in §1.2.2$_{36}$ we used marginal gambles to construct a set of marginally desirable gambles corresponding to a given lower prevision. We are going to do something very similar for conditional previsions: with each of the updated lower previsions present in a conditional prevision we can associate a set of marginally desirable (marginal) gambles, by patching these together as we did with the gambles in each of the $\mathscr{K}_B$ to obtain $\mathscr{K}$, we obtain a set of marginal gambles for the conditional prevision. So what we need to do is generalize definition (1.20)$_{37}$ to conditional previsions: Given a conditional lower prevision $\underline{P}(\bullet|\mathscr{B}):\mathscr{K}^\star \to \mathscr{L}_\Omega$, the marginal gambles are generated by the function $G_{\underline{P}(\bullet|\mathscr{B})}:\mathscr{K} \to (\mathscr{L}_\Omega)_{\mathscr{B}\text{-msr}}$, defined for every gamble f in $\mathscr{K}$ by

$$G_{\underline{P}(\bullet|\mathscr{B})} f = f - \underline{P}(f|\mathscr{B}). \tag{1.62}$$

Illustration with $\Omega := A \cup B$, $\mathscr{B} := \{A, B\}$, and $C := \{a, b\}$, the set we project on for visualization (with $a \in A$, $b \in B$):



The marginal gambles resulting from this generalized definition are marginally desirable:

(i) conceptually, because whichever event B in the partition $\mathscr{B}$ occurs, the restricted gamble $(G_{\underline{P}(\bullet|\mathscr{B})}f)_B$ is a marginally desirable gamble, as it is identical to the marginal gamble $G_{\underline{P}(\bullet|B)}f_B$;

(ii) mathematically, because it is identical to the finite sum

$$\sum_{B:\mathscr{B}}\left(G_{\underline{P}(\bullet|B)}f_B\right)_\Omega = \sum_{B:\mathscr{B}}\left(f_B - \underline{P}(f|B)\right)_\Omega = f - \sum_{B:\mathscr{B}}\underline{P}(f|B)\cdot I^B$$

of marginally desirable gambles (note the use of trivial extension to go from B to Ω).

These marginally desirable gambles form the set

$$\mathscr{G}_{\underline{P}(\bullet|\mathscr{B})} := \{G_{\underline{P}(\bullet|\mathscr{B})} f \mid f : \mathscr{K}\}. \tag{1.63}$$

Our basis is now strong enough to discuss, in the next two subsections,

(i) how to derive updated previsions from unconditional ones, and
(ii) the rationality criteria we need when working with both conditional and unconditional previsions.

1.3.3 *Natural & regular extension to updated previsions*

In the previous section, when we introduced natural extension or some rationality criterion for previsions, we did so by translating things to the language of sets of desirable gambles, then using its simple rules (cf. §1.2.1$_{34}$), and afterwards going back to the language of previsions. Here, we are going to do the same to find out how to extend an uncertainty model to updated previsions.

But first we need to introduce, for arbitrary sets of gambles $\mathscr{F} :\subseteq \mathscr{L}_\Omega$ and for any event C of the possibility space Ω, the contingent set of gambles

The support function: $\operatorname{supp} f = \{\omega : \Omega \mid f\omega \neq 0\}$.

$$\mathscr{F}_C := \{f_C \mid f : \mathscr{F} \wedge f \cdot I^C = f\} = \{f_C \mid f : \mathscr{F}_{\operatorname{supp} \bullet \subseteq C}\}, \tag{1.64}$$

where $\mathscr{F}_{\operatorname{supp} \bullet \subseteq C}$ is the set of gambles that reduce to the zero gamble on $\Omega \setminus C$.

With the definition (1.62)$_\cap$ of marginal gambles for conditional previsions we have a tool to do the first translation. Going back from some set of desirable gambles $\mathscr{R} :\subseteq \mathscr{L}_\Omega$ to an unconditional prevision is done with (1.12)$_{36}$ or one of its variants. To obtain an updated prevision for some conditioning event C of Ω, we apply the same formula, but, instead of $\mathscr{R}$, use its restriction to C [Walley 1991, §F2$_{615}$]: (let g be a gamble on C)

$$\begin{aligned} \underline{P}(g|C) &:= \sup\{\alpha : \mathbb{R} \mid I^C \cdot (g - \alpha)_\Omega \in \mathscr{R}\} \\ &= \sup\{\alpha : \mathbb{R} \mid g - \alpha \in \mathscr{R}_C\}, \end{aligned} \tag{1.65}$$

Here g_Ω in $\mathscr{L}_\Omega$ is the trivial extension of g from C to Ω, i.e., $(g_\Omega)_C = g$ and $(g_\Omega)_{\Omega\setminus C} = 0$ and $\mathscr{R}_C$ is the set of desirable gambles contingent on C.

When we start from some lower prevision $\underline{P}$ on $\mathscr{K}$, with $\mathscr{K} :\subseteq \mathscr{L}_\Omega$, formula (1.65) becomes the formula for natural extension to an updated prevision:

$$\operatorname{lce}_{\underline{P}}(g|C) := \sup\{\alpha : \mathbb{R} \mid g - \alpha \in (\mathscr{R}_{\underline{P}})_C\}, \tag{1.66}$$

which generalizes (1.23)$_{39}$. We have not done the effort to give the back-expanded version, as the generalized Bayes's rule we encounter in the next subsection provides an approach to conditioning that is usually more convenient.

Coherence of the updated prevision $\underline{P}(\bullet|C)$ is then a consequence of coherence of $\mathscr{R}_C$ or $(\mathscr{R}_{\underline{P}})_C$. As the expressions (1.65) and (1.66) defin-

ing the updated prevision $\underline{P}(\bullet|C)$ are formally identical to the relationship $(1.12)_{36}$ we had for unconditional previsions, $\underline{P}(\bullet|C)$ avoids sure loss and is coherent if it respectively satisfies $(1.25)_{40}$ and $(1.28)_{41}$. This must not surprise us: we introduced updated lower previsions at the beginning of §$1.3.2_{54}$ as normal lower previsions with some extra meaning added.

Before fully focusing our attention to coherence in the next subsection, it is interesting to realize the significance of (1.64) applied to sets of desirable gambles: It tells us how to obtain an updated uncertainty model in a way that takes borderline behavior into account. This is important for natural extension to updated previsions: they can depend critically upon borderline behavior.

To see this, consider a coherent set of desirable gambles $\mathscr{R} :\subset \mathscr{L}_\Omega$, the corresponding unconditional coherent lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$, and a conditioning event $C :\subset \Omega$ such that

$$\underline{P}C := \sup\{\alpha : \mathbb{R} \mid I^C - \alpha \in \mathscr{R}\} = 0.$$

This one assessment has grave consequences for $\mathscr{R}_C$ and $(\mathscr{R}_{\underline{P}})_C$. To find out which, consider any gamble f in $\mathscr{R}_{\text{supp}\bullet\subseteq C}$. By definition, $f \in \mathscr{R}$ and thus $\underline{P}f \geq 0$. The desirability of f also implies that $\sup\{f\} \geq 0$ (otherwise it would lead to a partial loss, cf. $(1.7)_{35}$), so $f \leq \sup\{f\} \cdot I^C$, from which $\underline{P}f \leq \sup f \cdot \underline{P}C = 0$ follows by monotonicity $(1.33)_{42}$. Combining these two constraints on $\underline{P}f$, we find that $\underline{P}f = 0$, and thus that $G_{\underline{P}}f = f$. This means that:

> Any desirable gamble that has as its support an event of lower probability zero is *marginally* desirable.

Formally, $\mathscr{R}_{\text{supp}\bullet\subseteq C} \subseteq \mathscr{G}_{\mathscr{R}}$ or $\mathscr{R}_C \subseteq (\mathscr{G}_{\mathscr{R}})_C$. As $f \notin \mathscr{G}_{\mathscr{R}}$ implies $\underline{P}f \neq 0$ (cf. $(1.18)_{37}$), $\mathscr{G}_{\underline{P}} = \mathscr{G}_{\mathscr{R}}$. Together with equation $(1.21)_{37}$, the last equality implies that $(\mathscr{D}_{\underline{P}})_C = \emptyset$, of which $(\mathscr{R}_{\underline{P}})_C = (\mathscr{L}_C)_{\geq 0 \wedge \neq 0}$ is the natural extension $(1.10)_{35}$: the updated models $(\mathscr{R}_{\underline{P}})_C$ we derived from $\underline{P}$ and the corresponding updated prevision $\underline{P}(\bullet|C)$ on $\mathscr{L}_C$ are vacuous.

Now consider the updated lower prevision $\underline{Q}(\bullet|C)$ on $\mathscr{L}_C$ corresponding to $\mathscr{R}_C$ and consider a desirable gamble g in $\mathscr{R}_C$, nontrivial in the sense that $\inf\{g\} \leq 0$, for which

$$\underline{Q}(g|C) := \sup\{\alpha : \mathbb{R} \mid g - \alpha \in \mathscr{R}_C\} > 0.$$

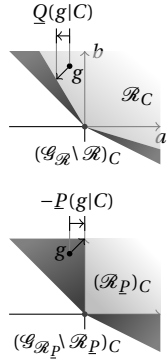
From the above derivation, we know that

$$\underline{P}(g|C) = \inf\{g\} \leq 0 < \underline{Q}(g|C). \tag{1.67}$$

This illustrates the fact that crucial (borderline) information about $\mathscr{R}_C$ is lost in the transition from $\mathscr{R}$ to $\underline{P}$ when $\underline{P}C = 0$.

Of course, if we initially have a given set of desirable gambles $\mathscr{R}$, the above is just a warning that we might be throwing away useful information when we only use $\mathscr{R}$ through the corresponding unconditional

Let $C := \{a, b\} \subset \Omega$, we visualize the part of the border of $\mathscr{R}$ lying in $(\mathscr{L}_\Omega)_{\text{supp}\bullet\subseteq C}$:



prevision $\underline{P}$. However, it also points out that previsions lack the modeling power to express

(i) some beliefs contingent on the assumption that some event happens, and at the same time

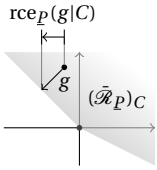
(ii) that this event is considered extremely unlikely to happen relative to the union of all the other – individually possibly equally unlikely – events.

It is only human to find this a bit limiting. What would we like to be different in our example leading up to $(1.67)_{\frown}$? We would like $(\mathscr{R}_{\underline{P}})_C$ to *not* be vacuous just because $\underline{P}C = 0$. This can be achieved by modifying either how we go from $\mathscr{G}_{\underline{P}}$ to $\mathscr{D}_{\underline{P}}$, i.e., $(1.21)_{37}$, or how we go from $\mathscr{D}_{\underline{P}}$ to $\mathscr{R}_{\underline{P}}$, i.e., natural extension $(1.10)_{35}$. The first path is intriguing but untrodden in the literature, so we here take the second, more familiar path of replacing natural extension by regular extension [Walley 1991, §J$_{639}$]. However, later on in §3.1.1$_{95}$, we will fruitfully explore the first path.

Given a lower prevision $\underline{P}$ on $\mathscr{K}$, regular extension takes the set of desirable gambles $\mathscr{R}_{\underline{P}}$ obtained by natural extension of $\mathscr{R}_{\underline{P}}$, but additionally considers any of its marginally desirable gambles $\mathscr{G}_{\mathscr{R}_{\underline{P}}}$ that has a strictly positive upper prevision to be desirable as well. We write the resulting set of desirable gambles [Walley 1991, §F$_{4615}$] with an additional bar on top to distinguish it from $\mathscr{R}_{\underline{P}}$:

$$\bar{\mathscr{R}}_{\underline{P}} := \mathscr{R}_{\underline{P}} \cup \{f : \mathscr{G}_{\mathscr{R}_{\underline{P}}} \mid \operatorname{mce}_{\underline{P}} f > 0\}, \tag{1.68}$$

where we have used the relationship $-\operatorname{lce}_{\underline{P}}(-f) = \operatorname{mce}_{\underline{P}} f$ to express the natural extension to upper previsions.



Now consider the conditioning event C again. The regular extension is defined, for every gamble g on C, by a modified version of $(1.66)_{56}$: the regular extension function

$$\operatorname{rce}_{\underline{P}}(g|C) := \sup\{\alpha : \mathbb{R} \mid g - \alpha \in (\bar{\mathscr{R}}_{\underline{P}})_C\} \tag{1.69}$$

Walley [1991, §J1$_{639}$] gives another, practically very useful form:

$$\operatorname{rce}_{\underline{P}}(g|C) = \begin{cases} \inf_{P : \mathscr{M}\underline{P} \wedge PC > 0} \frac{1}{PC} \cdot Pg_\Omega, & \operatorname{mce}_{\underline{P}} I^C > 0, \\ \inf\{g\}, & \text{otherwise.} \end{cases} \tag{1.70}$$

As we never (directly) condition on infinite partitions, joint coherence (cf. §1.3.4) is ensured [Miranda 2008*b*; Walley 1991, J$_{3640}$].

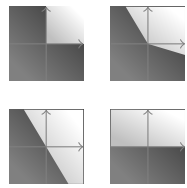
Formula (1.70) echoes both Bayes's theorem and $(1.55)_{50}$, which showed that the natural extension can be calculated as a lower envelope. An infimum is used because $\{P : \mathscr{M}\underline{P} \mid PC > 0\}$ need not be closed. When $\operatorname{ext}(\mathscr{M}\underline{P})$ is finite, it may replace $\mathscr{M}\underline{P}$ in this formula. (To see this, write the P in $\mathscr{M}\underline{P}$ for which $PC > 0$ as a convex combination of the elements of $\operatorname{ext}(\mathscr{M}\underline{P})$.)

When we use regular extension in this thesis, we explicitly mention it.

The natural and regular extension are not the two only possible updated previsions: every coherent prevision dominating the natural exten-

sion and dominated by the regular extension can be used as an updated prevision; an example is $\underline{P}(\bullet|C)$. In terms of desirable gambles, the difference with regular extension is that in (1.68) not all marginally desirable gambles with a positive upper prevision are chosen to be desirable. Care should be taken that when calculating updated previsions for multiple (overlapping) conditioning events, they should all derive from the *same* modified set of desirable gambles.

Regular and natural extension coincide for extension to unconditional previsions, since then $\underline{P}\Omega = 1$. Remark that when $\text{mce}_{\underline{P}}\, I^C = 0$, regular and natural extension to the corresponding updated prevision also coincide for the given prevision $\underline{P}$: they are both vacuous. However, in this case the whole of $\mathscr{L}_C$ is also marginally desirable. So now *any* coherent lower prevision on $\mathscr{L}_C$ can be taken as the updated prevision, because only coherence limits our freedom of choosing which marginally desirable gambles we take to be desirable.



1.3.4 *Separate coherence, joint coherence & the generalized Bayes's rule*

Now that we know how to relate unconditional, updated, and conditional previsions to sets of desirable gambles and vice versa, it is time to reflect on what rationality criteria we have to impose when we want to consistently combine the different types of previsions. Each of the previsions in such a consistent combination should be derivable by natural extension – or regular extension, if preferred – from the union of the sets of desirable gambles corresponding to each of these previsions. So combining is done by taking unions of sets of desirable gambles and consistency is guaranteed by invariance under natural extension.

The first type of combination, one we already encountered, is a conditional prevision $\underline{P}(\bullet|\mathscr{B}) : \mathscr{K}^\star \to \mathscr{L}_\Omega$, where $\mathscr{B} :\subseteq \wp\Omega$ is some partition and $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ some set of gambles. It can be useful to talk about the coherence of each of its separate, constitutive updated previsions at the same time. So therefore we call a conditional prevision with coherent constitutive updated previsions separately coherent [Walley 1991, §6.2$_{289}$].

Three interesting consequences of separate coherence are generalizations of properties (1.30)$_{42}$, (1.34)$_{42}$, (1.35)$_{42}$ for coherent lower previsions for which we can replace constants by measurable functions: (let $f : \mathscr{K}$, $g : (\mathscr{L}_\Omega)_{\mathscr{B}\text{-msr}\wedge\geq 0}$, $h : (\mathscr{L}_\Omega)_{\mathscr{B}\text{-msr}}$)

Walley [1991, §6.2.6$_{292}$] mentions a large number of consequences of separate coherence.

Measurable nonnegative homogeneity:

$$g \cdot f \in \mathscr{K} \;\Rightarrow\; \underline{P}(g \cdot f|\mathscr{B}) = g \cdot \underline{P}(f|\mathscr{B}), \tag{1.71}$$

Meas. normedness:

$$h \in \mathscr{K} \;\Rightarrow\; \underline{P}(h|\mathscr{B}) = h, \tag{1.72}$$

Meas. additivity:

$$f + h \in \mathscr{K} \;\Rightarrow\; \underline{P}(f + h|\mathscr{B}) = \underline{P}(f|\mathscr{B}) + h. \tag{1.73}$$

More challenging than only considering a conditional or uncondi-

tional prevision by itself, is to look at the joint coherence of an unconditional and one or more conditional previsions [Walley 1991, §6.3.3$_{293}$]. To get an idea what joint coherence is about, consider an unconditional prevision $\underline{P}$ on $\mathscr{K}$ and one conditional prevision $\underline{P}(\bullet|\mathscr{B}) : \tilde{\mathscr{K}}^\star \to \mathscr{L}_\Omega$, where $\mathscr{B} :\subset \wp\Omega$ is some partition of Ω and $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ and $\tilde{\mathscr{K}} :\subseteq (\mathscr{L}_\Omega)_{\mathscr{B}\text{-msr}}$ are two sets of gambles.

(i) With each prevision, there corresponds a set of marginal gambles, respectively $\mathscr{G}_{\underline{P}}$ and $\mathscr{G}_{\underline{P}(\bullet|\mathscr{B})}$, and thus

(ii) a set of desirable gambles, respectively $\mathscr{D}_{\underline{P}}$ and $\mathscr{D}_{\underline{P}(\bullet|\mathscr{B})}$.

(iii) Combining the two previsions consists in combining their corresponding sets of desirable gambles, i.e., this consists in forming the union $\mathscr{D}_{\underline{P}} \cup \mathscr{D}_{\underline{P}(\bullet|\mathscr{B})}$. (In this example, we could have taken the union after the first step, but this way of attacking things would also allow adding some set of desirable gambles to the mix.)

(iv) To check the consistency of this combination, we first take the natural extension for desirable gambles (1.10)$_{35}$:

$$\mathscr{R}_{\underline{P},\underline{P}(\bullet|\mathscr{B})} := \mathscr{R}_{\mathscr{D}_{\underline{P}} \cup \mathscr{D}_{\underline{P}(\bullet|\mathscr{B})}}. \tag{1.74}$$

(v) Then we use this set to define natural extension to unconditional and updated previsions (cf. (1.12)$_{36}$ and (1.66)$_{56}$): (let $f : \mathscr{L}_\Omega$, $C :\subset \Omega$, and $g : \mathscr{L}_C$)

$$\mathrm{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})} f := \sup\{\alpha : \mathbb{R} \mid f - \alpha \in \mathscr{R}_{\underline{P},\underline{P}(\bullet|\mathscr{B})}\}, \tag{1.75}$$

$$\mathrm{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})}(g|C) := \sup\{\alpha : \mathbb{R} \mid g - \alpha \in (\mathscr{R}_{\underline{P},\underline{P}(\bullet|\mathscr{B})})_C\}. \tag{1.76}$$

(vi) And finally – assuming that $\mathscr{R}_{\underline{P},\underline{P}(\bullet|\mathscr{B})}$ and $(\mathscr{R}_{\underline{P},\underline{P}(\bullet|\mathscr{B})})_B$ avoid sure loss (for all $B : \mathscr{B}$) – we can express joint coherence as invariance under natural extension (cf. §1.2.3$_{38}$): It must hold for all $f : \mathscr{K}$, $h : \tilde{\mathscr{K}}$, and $B : \mathscr{B}$ that

$$\underline{P} f = \mathrm{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})} f, \tag{1.77}$$

$$\underline{P}(h|B) = \mathrm{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})}(h_B|B). \tag{1.78}$$

Step (iii) shows that generalizing (1.74)–(1.78) to any finite number of non-conflicting uncertainty models is conceptually straightforward: taking unions of sets of desirable gambles is the key idea.

Now assume $\underline{P}$ and $\underline{P}(\bullet|\mathscr{B})$ to be separately and jointly coherent. The consequence of joint coherence that interests us most is similar to the following immediate consequences of (1.77) and (1.78): (now let $g : \tilde{\mathscr{K}}_B$)

Walley [1991, §6.3.5$_{296}$] mentions a large number of consequences of joint coherence.

$$\text{Cancellation:} \qquad G_{\underline{P}} f \in \mathscr{K} \quad \Rightarrow \quad \underline{P}(G_{\underline{P}} f) = 0, \tag{1.79}$$

$$\text{Updated cancellation:} \qquad G_{\underline{P}(\bullet|B)} g \in \tilde{\mathscr{K}}_B \;\Rightarrow\; \underline{P}(G_{\underline{P}(\bullet|B)} g|B) = 0. \tag{1.80}$$

It expresses that an updated marginal gamble is in some sense also a plain marginal gamble:

$$\text{Mixed cancellation:} \quad (G_{\underline{P}(\bullet|B)} g)_\Omega \in \mathscr{K} \quad \Rightarrow \quad \underline{P}(G_{\underline{P}(\bullet|B)} g)_\Omega = 0. \tag{1.81}$$

Formula (1.81) expresses a particularly interesting relationship between unconditional and updated previsions: Using the definition of a marginal gamble $(1.20)_{37}$, (1.77), and (1.78), we can rewrite the implicand as $\operatorname{lce}_{\underline{P},\underline{P}(\bullet|B)}\big(g-\operatorname{lce}_{\underline{P},\underline{P}(\bullet|B)}(g|B)\big)_{\Omega}=0$, which holds for any g in $\mathscr{L}_B$. Furthermore, it remains valid for any set of uncertainty models we are extending; so consider that we only started out with the unconditional prevision $\underline{P}$, then the expression becomes $\operatorname{lce}_{\underline{P}}\big(g-\operatorname{lce}_{\underline{P}}(g|B)\big)_{\Omega}=0$. This is the basis for the generalized Bayes's rule or GBR [Walley 1991, §6.4.1_{297}], an alternative to $(1.66)_{56}$ for calculating the natural extension to an updated prevision:

$$\begin{aligned}&\operatorname{lce}_{\underline{P}} I^B>0 \;\Rightarrow\; \operatorname{lce}_{\underline{P}}(g|B)=\mu,\\ &\qquad\text{where } \mu:\mathbb{R} \text{ is the unique solution of } \operatorname{lce}_{\underline{P}}(g-\mu)_{\Omega}=0.\end{aligned} \tag{1.82}$$

Checking whether the conditioning event has a positive lower probability or not is necessary. In the latter case, the solution is not unique; however, we have already seen in §1.3.3_{56} that the natural extension then becomes vacuous.

Formula (1.82) reduces to Bayes's rule when it is applied to linear previsions, whence its name. This fact can be used in combination with $(1.55)_{50}$ and the monotone character of $P:\mathscr{M}\underline{P};\frac{1}{PB}\cdot Pg_{\Omega}$ to obtain a lower envelope theorem for updated previsions [Walley 1991, §6.4.2_{298}]:

$$\operatorname{lce}_{\underline{P}}(g|B)=\begin{cases}\min_{P:\operatorname{ext}(\mathscr{M}\underline{P})}\frac{1}{PB}\cdot Pg_{\Omega}, & \operatorname{lce}_{\underline{P}} I^B>0,\\ \inf\{g\}, & \text{otherwise.}\end{cases} \tag{1.83}$$

It is interesting to compare this to the formula $(1.70)_{58}$ for regular extension.

Up to this point in this section, we have had an overview of the important basic ideas concerning conditional and updated previsions. The remaining subsections briefly state and contextualize related results that we use further on in this thesis.

1.3.5 *Marginal extension*

In the last subsection, we discussed how to combine an unconditional prevision $\underline{P}$ on $\mathscr{K}$ with a conditional prevision $\underline{P}(\bullet|\mathscr{B})$ in $\tilde{\mathscr{K}}^{\star}\to\mathscr{L}_{\Omega}$, where $\mathscr{B}$ is a partition of Ω and where $\mathscr{K}\coloneqq\subseteq\mathscr{L}_{\Omega}$ and $\tilde{\mathscr{K}}\coloneqq\subseteq(\mathscr{L}_{\Omega})_{\mathscr{B}\text{-msr}}$ are two sets of gambles.

Whenever $\mathscr{K}$ only contains $\mathscr{B}$-measurable gambles, $\underline{P}$ only contains information about the relative likelihoods of the different events of the partition $\mathscr{B}$ and none about the relative likelihoods within these events. Therefore, $\underline{P}$ can be seen as a marginal prevision (cf. second-to-last paragraph of §1.3.1_{52}). On the other hand, $\underline{P}(\bullet|\mathscr{B})$ by definition can only encode relative likelihoods within the events of the partition $\mathscr{B}$.

This hierarchical compartmentalization of information greatly simplifies the calculation of the natural extension. Assume $\underline{P}$ and $\underline{P}(\bullet|\mathscr{B})$ avoid sure loss. After extending $\underline{P}$ to a coherent prevision on $(\mathscr{L}_\Omega)_{\mathscr{B}\text{-msr}}$ and $\underline{P}(\bullet|\mathscr{B})$ to a separately coherent conditional prevision on $\mathscr{L}_\Omega$, we get the combined model's natural extension for free with Walley's marginal extension theorem [1991, §6.7.2$_{314}$]: (let f be a gamble on Ω)

$$\operatorname{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})} f := \operatorname{lce}_{\underline{P}}\big(\operatorname{slce}_{\underline{P}(\bullet|\mathscr{B})}(f|\mathscr{B})\big), \tag{1.84}$$

where we have used separate natural extension

$$\operatorname{slce}_{\underline{P}(\bullet|\mathscr{B})}(f|\mathscr{B}) := \textstyle\sum_{B:\mathscr{B}} \operatorname{lce}_{\underline{P}(\bullet|B)}(f_B|B)\cdot I^B. \tag{1.85}$$

of the updated previsions specified by $\underline{P}(\bullet|\mathscr{B})$, whose definition echoes the definition (1.60)$_{54}$ of a conditional lower prevision.

Note that the marginal extension theorem does not hold when $\mathscr{K}$ contains other than only $\mathscr{B}$-measurable gambles. In such a situation, the right hand side of (1.84) will be dominated by the natural extension, as it erases any information available in the previsions for these non-$\mathscr{B}$-measurable gambles.

The specific case where the possibility space is a cartesian product of finite sets and the lower previsions are given as lower envelopes of sets of linear previsions deserves some more attention, as it will be useful in §3.2.5$_{131}$: Let $\Omega := \mathscr{X}\times\mathscr{Y}$, where $\mathscr{X}$ and $\mathscr{Y}$ are finite sets, let $\mathscr{K} := \mathscr{L}_{\mathscr{X}}$ and $\tilde{\mathscr{K}} := \mathscr{L}_{\mathscr{Y}}$, and let $\mathscr{B} := \{\imath x \mid x:\mathscr{X}\}$. Moreover, let $\underline{P}g := \inf_{\gamma:\Gamma} P_\gamma g$ and $\underline{P}(h|\imath x) := \inf_{\zeta_x:Z_x} P_{\zeta_x} h$, where g and h can be any gamble on $\mathscr{X}$ and $\mathscr{Y}$, respectively, and where Γ and $Z := \bigtimes_{x:\mathscr{X}} Z_x$ are index sets. Then for every gamble f on Ω the marginal extension can be written as

The elastic version of the Cartesian product operator $\times$ is $\bigtimes$; e.g., $\bigtimes_{\ell:1..3}\mathbb{R}^\ell = \mathbb{R}^1\times\mathbb{R}^2\times\mathbb{R}^3$.

$$\begin{aligned}\operatorname{lce}_{\underline{P},\underline{P}(\bullet|\mathscr{B})} f &= \underline{P}\big(\underline{P}(f|\mathscr{B})\big)\\ &= \inf_{\gamma:\Gamma} P_\gamma(\inf_{\zeta_\bullet:Z_\bullet} P_{\zeta_\bullet} f)\\ &= \inf_{\gamma:\Gamma}\inf_{\zeta:Z} P_\gamma(P_{\zeta_\bullet} f)\\ &= \inf_{\gamma:\Gamma}\inf_{\zeta:Z} P_\gamma\Big(\textstyle\sum_{x:\mathscr{X}} I^x\cdot P_{\zeta_x}\big(f(x,\bullet)\big)\Big),\end{aligned} \tag{1.86}$$

where the next to last step – bringing the inner infimum outside – has been proven by Miranda & De Cooman [2007, §4$_{208}$] and the last step is meant to make sure the notation is clear.

1.3.6 *Independent products*

Our aim in this subsection is to combine a finite tuple $(\underline{P}_j \mid j:J)$ of given coherent lower previsions from $\bigtimes_{j:J}\mathscr{P}\mathscr{K}_j$; here, J is some finite index set and for all j in J, $\mathscr{K}_j :\subseteq \mathscr{L}_{\Omega_j}$, with Ω_j some possibility space. These previsions have to be combined into a coherent joint lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$, where $\Omega := \bigtimes_{j:J}\Omega_j$, of which the Ω_j-marginals restricted to $\mathscr{K}_j$ coincide with $\underline{P}_j$.

Without additional information or assumptions about the relationships between the beliefs expressed by the marginals, this combination can be done by first defining $\underline{P}$ on the set of cylindrically extended gambles (by letting $\underline{P}\tilde{f} = \underline{P}_j f$, for all j in J and all gambles f on Ω_j) and then taking the natural extension $(1.23)_{39}$ to the whole of $\mathscr{L}_\Omega$.

However, we here restrict ourselves to the assumption that the beliefs about the different components of Ω are independent. To be more precise: for any proper subset I of J, the beliefs about $\Omega_I := \bigtimes_{i:I}\Omega_i$ are independent of the beliefs about $\Omega_K := \bigtimes_{k:K}\Omega_k$, where K is any subset of J that is disjoint from I. Without further assumptions, this independence is formalized by saying that $\underline{P}\big(\tilde{f} \,\big|\, \{\imath\omega \times \Omega_I \,|\, \omega : \Omega_K\}\big)$ must be equal to $\underline{P}_i f$ for all i in I and all gambles f on Ω_i [Walley 1991, §9.3.1_{452}].

Once we have these partially specified conditional previsions, the least committal independent joint lower prevision $\underline{P}$ that is coherent with them can – if it exists – be obtained using the so-called independent natural extension procedure [Walley 1991, §9.3.2_{452}]. Conceptually, this procedure corresponds to applying steps (i)–(v)$_{60}$ in §1.3.4_{59} to the set of partially specified conditional previsions. An independent joint prevision is also called a product prevision.

Usually, the independent natural extension of a tuple of marginals is not the only coherent product prevision. In fact, it can be quite hard to calculate, so other products – although not least committal – may become interesting from a computational point of view. Some of these other products become a natural choice after making extra assumptions, for example by using a more restrictive interpretation of the theory of imprecise probabilities.

The interpretation we use to lead us to a computationally simpler product is the so-called sensitivity analysis interpretation (see [Walley 1991, §1.1.5_6,§2.10.4_{105},§5.9_{253}] and also §3.3.2_{140}). Under this interpretation, the independence assumption is formalized by the requirement that the product is an independent lower envelope.

To define such an independent lower envelope, we need to introduce independent products of linear previsions. First consider two linear previsions, Q on $\mathscr{L}_\mathscr{X}$ and R on $\mathscr{Y}$; their independent product, denoted $Q \times R$, is defined for every gamble h on $\mathscr{X} \times \mathscr{Y}$ by

$$(Q \times R)h = Q(Rh). \tag{1.87}$$

The corresponding elastic operator is denoted by $\bigtimes$. As in general the product order matters, $\bigtimes$ is defined for tuples and not sets of linear previsions. However, in this thesis we only encounter products of σ-additive linear previsions, for which the order does not matter, as the Fubini theorem can be applied [Burrill 1972, §7-6_{125}]. This does imply however that in these cases we must restrict attention to measurable gambles.

Elastic operators are defined recursively; e.g., if $(P_j \,|\, j : 1..7)$ is a tuple of linear previsions, then $\bigtimes_{j:1..7} P_j = P_1 \times \bigtimes_{j:2..7} P_j$.

An independent product $\underline{P}$ is an independent lower envelope when

there is a set of linear previsions $\mathscr{Q} :\subseteq (\mathscr{P}\mathscr{L}_\Omega)_{\text{lin}}$ such that $\underline{P} f = \inf_{P:\mathscr{Q}} P f$ for any measurable gamble f on Ω and such that each P in $\mathscr{Q}$ has independent Ω_j-marginals Q_j, i.e., $P = \bigtimes_{j:J} Q_j$ [Walley 1991, §9.1.5$_{446}$]. By gathering these marginals in sets of linear previsions $\mathscr{Q}_j :\subseteq (\mathscr{P}\mathscr{L}_{\Omega_j})_{\text{lin}}$, we can express that $\underline{P}$ is an independent lower envelope of the marginals $(\underline{P}_j \mid j : J)$ by requiring that $\underline{P}_j g_j = \inf_{Q_j:\mathscr{Q}_j} Q_j g_j$ for all j in J and measurable gambles g_j on Ω_j.

Now, when each marginal $\underline{P}_j$ is *specified* as a lower envelope of a set of linear previsions $\mathscr{Q}_j$, the natural thing to do is define

$$\mathscr{Q} := \{\bigtimes_{j:J} Q_j \mid Q : \bigtimes_{j:J} \mathscr{Q}_j\}$$

and consequently, for all suitably measurable gambles f,

$$\underline{P} f := \inf_{P:\mathscr{Q}} P f = \inf\{(\bigtimes_{j:J} Q_j) f \mid Q : \bigtimes_{j:J} \mathscr{Q}_j\} \tag{1.88}$$

$$= \min\{(\bigtimes_{j:J} Q_j) f \mid Q : \bigtimes_{j:J} \mathscr{M}\underline{P}_j\}, \tag{1.89}$$

Illustrating the convex hull and closure operators: $\text{co}\{0,1\} = [0,1]$, $\text{cl}]0,1[= [0,1]$.

where the equality follows from

$$\mathscr{Q} \subseteq \{\bigtimes_{j:J} Q_j \mid Q : \bigtimes_{j:J} \text{cl co}\, \mathscr{Q}_j\} = \{\bigtimes_{j:J} Q_j \mid Q : \bigtimes_{j:J} \mathscr{M}\underline{P}_j\} \subset \text{cl co}\, \mathscr{Q}$$

and the fact that taking the infimum over the convex closure does not change its value. An independent product defined by an independent lower envelope such as (1.89) is called a type-1 product [Walley 1991, §9.3.5$_{455}$]; the equivalent form (1.88) is the one we will encounter later on in some applications (cf. §3.4.4$_{149}$ and §4.3.2$_{182}$).

Having arrived at the end of this subsection, we have arrived at this foundation-laying chapter's end, ready for the things that this thesis has in store. Onwards!


SOIS
BEL
$e^{i\pi}=-1$

EXTREME LOWER PROBABILITIES

> Après dîner, Mlle de La Mole, loin de fuir Julien, lui parla et l'engagea en quelque sorte à la suivre au jardin; il obéit. Cette épreuve lui manquait. Mathilde cédait, sans trop s'en douter, à l'amour qu'elle reprenait pour lui. Elle trouvait un plaisir extrême à se promener à ses côtés; c'était avec curiosité qu'elle regardait ces mains qui, le matin, avaient saisi l'épée pour la tuer.
>
> Stendhal [1830, Ch. XVIII, ¶6]

In the previous chapter, we have talked at length about lower previsions. We have encountered the predicates asl, coh, and lin that characterize all those lower previsions that avoid sure loss, are coherent, or are linear. We have seen that, for lower previsions defined on finite possibility spaces, the set of all linear previsions could be represented by a unit simplex.

As a convex polytope, a unit simplex is fully characterized by its extreme points. These extreme points represent the degenerate previsions. Thus, the set of all linear previsions on some finite possibility space Ω is a polytope also, fully characterized by the degenerate previsions on Ω. They are the *extreme* linear previsions.

In this chapter, we obtain similar results when considering avoiding sure loss, coherence, and some other interesting properties. To wit, the sets of all lower probabilities satisfying one or more of these properties are convex polyhedra or polytopes; their extreme points are called extreme lower probabilities. We show the steps that have to be taken to compute these extreme lower probabilities starting from the expressions of predicates such as asl and coh (§2.1), fill in the technical details (§2.2$_{70}$), and give some practical results (§2.3$_{85}$). More extensive practical results can be found in the Herbarium$_{194}$.

Quaeghebeur & De Cooman [2006, 2008] give a less elaborate exposition than the present one.

Note that in this chapter we restrict ourselves to finite possibility spaces Ω and probabilities, i.e., lower previsions defined on the set of all indicators $\mathscr{I}_\Omega$ or, equivalently, on the power set $\wp\Omega$ of the possibility space. Recall that we only need to look at lower probabilities because of conjugacy; the extreme upper probabilities follow from the conjugacy relation for probabilities (1.15)$_{37}$.

2.1 Constraints & vertex enumeration

In this section, we introduce some concepts and results from polytope theory. We also show how polytope theory can be useful when studying

the different properties of interest that a lower probability can satisfy. In the end, this will allow us to formulate how to compute the extreme lower probabilities corresponding to each of these properties.

2.1.1 *Constraints*

Lower probabilities are defined on the power set $\wp\Omega$ and thus have $2^{|\Omega|}$ components, each indexed by one of the subsets of Ω. A lower probability is therefore a point in the $2^{|\Omega|}$-dimensional real vector space $\wp\Omega \to \mathbb{R}$.

We have seen in §1.2.9$_{50}$ that the set of all linear previsions is determined by a set of linear constraints (linear inequalities and linear equalities). A lot of interesting properties such as avoiding sure loss and coherence can also be expressed as a set of linear constraints. Actually, the definition of these properties usually consists of such a set of constraints.

We do not consider requirements related to independence and conditioning, which typically lead to nonlinear constraints.

As an example, let us look at the definition for avoiding sure loss: Consider a lower probability $\underline{P} : \mathscr{P}\mathscr{I}_\Omega$, then (1.25)$_{40}$ tells us that

$$\begin{aligned} \operatorname{asl}\underline{P} \Leftrightarrow \forall \mathscr{N} :\subseteq \mathscr{I}_\Omega;& \\ \forall \lambda : (\mathbb{R}_{>0})^{\mathscr{N}};& \\ \sup\{\textstyle\sum_{g:\mathscr{N}} \lambda_g \cdot G_{\underline{P}} g\} \geq 0.& \end{aligned} \tag{2.1}$$

We work with lower probabilities, so we can reformulate this definition in terms of events (i.e., subsets of Ω or, equivalently, elements of $\wp\Omega$). To be able to do this, we need to write out the marginal gambles, which is done using (1.20)$_{37}$. Also taking the finitary nature of Ω and therefore $\wp\Omega$ into account to further simplify the expression results in

$$\begin{aligned} \operatorname{asl}\underline{P} \Leftrightarrow \forall \lambda : (\mathbb{R}_{\geq 0})^{\wp\Omega};& \\ \textstyle\sum_{B:\subseteq\Omega} \lambda_B \cdot \underline{P}B \leq \sup\{\sum_{B:\subseteq\Omega} \lambda_B \cdot I^B\}.& \end{aligned} \tag{2.2}$$

This expression shows that the definition of avoiding sure loss consists of a set of linear constraints (linear inequalities) in $\underline{P}$, where each constraint is defined by some λ. To be precise, each of these constraints is a linear predicate on $\mathscr{P}\mathscr{I}_\Omega$.

Recall that lower probabilities are points in the vector space $\wp\Omega \to \mathbb{R}$; this means that linear constraints correspond to hyperplanes (linear equalities) or half-spaces defined by some hyperplane (linear inequalities). For example, in the expression for the linear constraints in (2.2), the left-hand side scalar product $\sum \lambda \cdot \underline{P} = \sum_{B:\subseteq\Omega} \lambda_B \cdot \underline{P}B$ determines the orientation of the hyperplane and the right-hand side constant $\sup\{\sum_{B:\subseteq\Omega} \lambda_B \cdot I^B\}$ determines its location. The convex intersection of these linear subspaces of $\wp\Omega \to \mathbb{R}$ is the set of lower probabilities that satisfies the property of interest (which in the above example is avoiding sure loss).

Generalizing the above example, a linear inequality constraint for a lower probability $\underline{P}$ on $\wp\Omega$ has the following form: (let $\lambda : \mathbb{R}^{\wp\Omega}$ be a vector of coefficients and let α be some real constant)

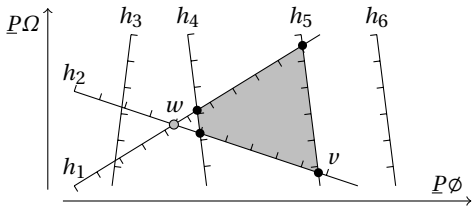
$$\sum \lambda \cdot \underline{P} \geq \alpha. \tag{2.3}$$

A linear equality constraint corresponds to two inequality constraints.

To develop more geometric intuition, the next subsection contains a graphical illustration of the concepts introduced above. Some new concepts related to linear constraints are also introduced and illustrated.

2.1.2 *A toy example*

Any graphical illustration is limited to two or three dimensions at most. As we are interested in depicting the space $\wp\Omega \to \mathbb{R}$ for some Ω, we must limit ourselves to a possibility space with one element. Its power set consist of only two elements, Ω and $\emptyset$. This results in a very simple toy example, which is nevertheless very illustrative.



Consider a tuple $h \coloneqq (h_i \mid i : 1..6)$ of constraints; each of them corresponds to a linear inequality of the form (2.3), i.e.,

$$\lambda_\emptyset \cdot \underline{P}\emptyset + \lambda_\Omega \cdot \underline{P}\Omega \geq \alpha,$$

where the relative values of the real numbers $\lambda_\emptyset$, λ_Ω, and α are implicitly defined in the picture. They are represented using lines with hairs that indicate the half-space they correspond to. The convex intersection of the half-spaces – i.e., the set $(\mathscr{P}\mathscr{I}_\Omega)_{\forall h}$ of lower probabilities that satisfy the constraints – is colored gray.

Notice that there are some constraints that can be removed from h without changing the resulting intersection; these are called redundant constraints. In our example, constraints h_6 is redundant because h_5 is more stringent and h_3 is made redundant by the set $\{h_1, h_2, h_4\}$.

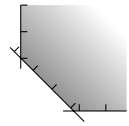
We can immediately identify the extreme lower probabilities corresponding to the given set of constraints: these are the extreme points (or vertices) of the gray area, indicated using a black dot $\bullet$ (v for example). All of them are formed by an intersection of nonredundant constraints. However, not every intersection of nonredundant constraints is an extreme lower probability (w for example, indicated using a little circle $\circ$).

What we have learned about constraints and vertices with this toy example can be extended to other, more realistic situations. The next subsection cites the results that underpin this generalization.

2.1.3 *Polyhedra, polytopes & vertex enumeration*

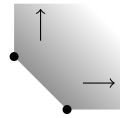
First, the intersection of a finite set of linear constraints, i.e., the points in $\wp\Omega \to \mathbb{R}$ satisfying the constraints, is a convex polyhedron by definition

(in algebraic geometry). A polytope is a bounded polyhedron. With each constraint there corresponds a so-called defining hyperplane. A simple illustration of an unbounded polyhedron is given on the side.



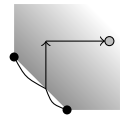
The Minkowski-Weyl theorem [Fukuda 2004] tells us that a convex polyhedron that does not include a linear subspace can be equivalently described by its

(i) vertices or extreme points: polyhedron points defined by the intersection of at least $2^{|\Omega|}$ defining hyperplanes (• in the illustration), and



(ii) extreme rays: (normalized) polyhedron directions defined by the intersection of $2^{|\Omega|} - 1$ defining hyperplanes and the hyperplane at infinity (⟶ in the illustration).

This means that any point of a convex polyhedron (◦ in the illustration) can be written as a not necessarily unique sum of



(i) a convex combination of extreme points (i.e., a decomposition in barycentric coordinates), and

(ii) a positive linear combination of extreme rays.

As polytopes are bounded, they can be described using extreme points only. Extreme points and extreme rays are the extreme lower probabilities we are looking for. In case the polyhedron does contain a linear subspace – e.g., if it is a half-space – additional points of the border of the linear subspace that take up the role of vertices – such as a point on the border of the half-space – need to be specified.

The above decomposition property is a reason why finding the extreme lower probabilities can be useful: Any lower probability satisfying some set of properties can be written in terms of the extreme lower probabilities corresponding to this set of properties. (Maaß [2003*a*,*b*] extends this result to a more general, infinitary context.) This may prove useful for theoretical as well as practical applications.

De Cooman & Miranda [2007, §9.1] and Miranda et al. [2008*a*, §5] provide examples of theoretical use.

One possible practical application is what actually led us to investigate extreme lower probabilities: We wanted to use them for efficient calculation or approximation of lower probabilities. The idea was that once the extreme lower probabilities have been determined, they can be used to calculate the other lower probabilities using the decomposition, or, by retaining only part of the decomposition, which can be used for approximation.

Anyway, the important point is that sets of lower probabilities can be equivalently defined either with sets of constraints or with sets of extreme lower probabilities. The former are typically available, due to the constraint-based nature of many rationality criteria in probability theory, the latter are not, but are judged by us to be potentially useful. Luckily, polytope theory provides techniques for vertex enumeration [Matheiss & Rubin 1980]: converting sets of linear constraints to sets of extreme points.

Ziegler [1995] gives an overview of polytope theory, Fukuda [2004] focuses on computation.

We used implementations by Avis [2000] and Fukuda & Prudon [1996]. They can deal with linear subspaces and also contain redundant constraint removal algorithms.

Freely available implementations of these vertex enumeration algorithms can be used in practice. The computational complexity of these algorithms is described in terms of three parameters: the dimension of the space, the number of (nonredundant) constraints and the number of vertices. It is an open problem whether there exists an algorithm that can enumerate the vertices in polynomial time and space (in the given parameters). The efficiency of these algorithms furthermore depends on the type of polytope; for example, whether or not it is simple [Avis et al. 1997].

For our application, the dimension $2^{|\Omega|}$ of the space grows exponentially with the size $|\Omega|$ of the possibility space. For most of the properties studied, this is accompanied by a corresponding observed increase in the number of constraints (see, e.g., the definition $(2.13)_{75}$ for monotonicity). Therefore, the time needed for vertex enumeration grows at least exponentially in the size of the possibility space. At the time of writing, computations (on a personal computer) are limited to possibility spaces up to around five elements.

So, even though it is not practical for nonsmall possibility spaces, the conclusion we wish to distill from this subsection is the following: Our quest for the extreme lower probabilities that satisfy some property, is – in principle – solved when the constraints corresponding to that property are known; the rest of the work is left to existing vertex enumeration algorithms.

2.2 PROPERTY CONSTRAINT GENERATION

Vertex-based definitions also exist; e.g., linear-vacuous mixtures are a convex combination of the degenerate previsions and the vacuous prevision [Walley 1991, §2.9.2$_{93}$].

We have seen that the search for extreme lower probabilities has been reduced to a search for constraints. This section contains constraint-based definitions for many of the properties that are of interest to people working with lower probabilities, such as $(2.2)_{67}$ for avoiding sure loss. Moreover, we obtain versions that are *feasible* in the sense that they generate a finite number of constraints (note that $(2.2)_{67}$ gives an infinite number). We also try to limit the number of redundant constraints generated, even though this is of less practical importance due to the existence of redundant constraint removal algorithms.

Once a feasible constraint-based definition is available for two or more properties, it is straightforward to obtain the extreme lower probabilities of the set of lower probabilities that satisfy all the properties at once; one just combines the generated list of constraints before using a vertex enumeration algorithm. We make use of this in our examples.

2.2.1 *Normedness, nonnegativity & additivity*

Classical probabilities are often thought of – next to betting rates – in terms of a unit mass distributed over the elementary events. This associ-

ation makes requirements placed on such classical probabilities almost tangible: they are normed, nonnegative, and additive.

Almost always, lower probabilities are also required to be normed and nonnegative. Even though we usually will not consider these requirements separately, but rather as a consequence of coherence, they are simple, so it is instructive to treat them here in their own right, i.e., give constraint-based definitions. For the sake of completeness, we complement this subsection with a constraint-based definition of additivity, a requirement that a lower probability does not need to satisfy.

For lower probabilities $\underline{P}$ on $\wp\Omega$, normedness (cf. $(1.34)_{42}$ for previsions) is formalized using a predicate $\mathrm{nrm}:\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$; it can be expressed – independently of the cardinality of Ω – using two linear equality constraints:

$$\mathrm{nrm}\,\underline{P} \Leftrightarrow \underline{P}\emptyset = 0 \wedge \underline{P}\Omega = 1. \tag{2.4}$$

Nonnegativity (cf. accepting sure gains $(1.29)_{42}$ for previsions) is formalized using a predicate $\mathrm{nng}:\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$; it can be expressed using $2^{|\Omega|}$ linear inequality constraints:

$$\mathrm{nng}\,\underline{P} \Leftrightarrow \underline{P} \geq 0. \tag{2.5}$$

Additivity is formalized using the predicate $\mathrm{add}:\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$ which could be expressed by giving, for every event A of Ω and every nontrivial partition $\mathscr{B} :\subseteq \wp A$ of this event, a linear equality constraint $\underline{P}A = \sum \underline{P}_{\mathscr{B}}$. Due to transitivity, it is enough to look at the partition consisting of all singletons. This allows us to eliminate a lot of redundant constraints and to end up with a definition using $2^{|\Omega|} - |\Omega|$ linear equalities:

$$\begin{aligned} \mathrm{add}\,\underline{P} \Leftrightarrow \forall A :\subseteq \Omega \mathbin{\mathbb{A}} |A| \neq 1;& \\ \underline{P}A = \textstyle\sum \underline{P}_A,& \end{aligned} \tag{2.6}$$

where, recall, $\underline{P}_A$ is the restriction of $\underline{P}$ to A. Note that $\underline{P}\emptyset = 0$ is included in this definition, as $\underline{P}_\emptyset$ has an empty domain, which implies $\sum \underline{P}_\emptyset = 0$.

Expressions such as (2.6) and the much more complex ones we shall encounter later are – once objects have been defined for sets and constraints – easy to translate to a computer program. The universal quantifier $\forall$ introduces a for-loop and the filter $\mathbb{A}$ an if-then-statement, which, if passed, adds the resulting constraint to the list of constraints for the property under scrutiny (here additivity).

2.2.2 *Superadditivity*

One of the big differences between lower probabilities and classical probabilities is the fact that – as a consequence of the typical rationality criteria used – the additivity requirement is relaxed to a superadditivity requirement.

Analogously to additivity, superadditivity is formalized using a predicate $\mathrm{sad} :\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$ which could be expressed by giving, for every event A of Ω and every partition $\mathscr{B} :\subseteq \wp A$ of this event, a linear inequality constraint $\underline{P}A \geq \sum \underline{P}_{\mathscr{B}}$. Due to transitivity, it is enough to look at the binary partitions. This allows us to eliminate a lot of redundant constraints:

$$\begin{aligned} \mathrm{sad}\,\underline{P} \Leftrightarrow \forall A :\subseteq \Omega;& \\ \forall B, C :\subseteq (\wp A)^2 \wedge B \cap C = \emptyset \wedge B \cup C = A;& \\ \underline{P}A \geq \underline{P}B + \underline{P}C.& \end{aligned} \tag{2.7}$$

However, superadditivity is not strong enough as (the mathematical formulation of) a rationality criterion: it does not imply avoiding sure loss. This can be quickly illustrated with the lower probability $\underline{P}$ in $(\mathscr{P}\mathscr{I}_{\{a,b,c\}})_{\mathrm{nrm}\wedge\mathrm{sad}}$ which is zero in singletons and one in doubletons; explicitly: take λ in $(2.2)_{67}$ to be one for doubletons and zero elsewhere, then

$$3 = \underline{P}\{a,b\} + \underline{P}\{a,c\} + \underline{P}\{b,c\} \not\leq \sup\{I^{\{a,b\}} + I^{\{a,c\}} + I^{\{b,c\}}\} = 2.$$

So how do we make sure that a superadditive lower probability avoids sure loss? The next subsection gives a type of approach we have not yet come across; the one after that treats avoiding sure loss itself.

2.2.3 *k-Monotonicity*

Thinking in terms of mass distribution, superadditivity corresponds to the idea that we do not have enough information to unequivocally distribute the whole of the unit mass over the elementary events and assigning the remains of the unit mass to the nonelementary events. The rationale behind this could be that the nature of the available information allows us to say to some extent how probable some nonelementary event is, without being able to be more specific.

Formalizing this idea leads to so-called belief functions, a special type of lower probabilities that can be written as convex combinations of vacuous probabilities relative to subsets (introduced in $(1.38)_{43}$). Each vacuous probability in the convex combination corresponds to a nonempty event; the coefficients correspond to the mass assigned to that event, elementary or not. Formally, let us characterize the belief functions with a predicate $\mathrm{bel} :\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$, which holds when a probability mass assignment can be found that generates the probability in question:

Dempster [1967] first introduced belief functions; Shafer [1976] based a theory of evidence on them.

For the infinite case, the extreme belief functions are those that are 1 on a filter of events and 0 elsewhere [Brüning & Dennenberg 2008].

$$\begin{aligned} \mathrm{bel}\,\underline{P} \Leftrightarrow \exists \mu : \Delta_{(\wp\Omega)_{\neq\emptyset}};& \\ \underline{P} = \textstyle\sum_{B:(\wp\Omega)_{\neq\emptyset}} \mu_B \cdot \underline{P}^B,& \end{aligned} \tag{2.8}$$

(Recall that we use subscripting to indicate coefficient vector components such as μ_B.) Note that this is a vertex-based definition, the vertices

– i.e., the extreme belief functions – being the vacuous probabilities relative to subsets.

Belief functions are not the only lower probabilities that can be written in terms of vacuous probabilities when we allow generalized mass assignments that do not sum up to one, have negative components, and a possible nonzero assignment for the empty set. Such a lower probability $\underline{P}$ on $\wp\Omega$ can be written in terms of its generalized mass assignment $\mu : \mathbb{R}^{\wp\Omega}$ and vice versa by Möbius inversion and Möbius transformation respectively [Shafer 1976, Lemma 2.3]:

$$\begin{aligned} &\forall A :\subseteq \Omega ; \underline{P}A = \textstyle\sum_{B:\subseteq A} \mu_B \\ \Leftrightarrow &\forall A :\subseteq \Omega ; \mu_A = \textstyle\sum_{B:\subseteq A} (-1)^{|A\setminus B|} \cdot \underline{P}B. \end{aligned} \tag{2.9}$$

There is a class of lower probabilities, the so-called k-monotone lower probabilities (where $k : \mathbb{N}_{>0}$), that can be described by specific types of generalized mass assignments [Chateauneuf & Jaffray 1989]. In general, these cannot be written as a convex combination of vacuous probabilities. We can find the extreme k-monotone lower probabilities, however, using a constraint-based definition of k-monotonicity. For every $k : \mathbb{N}_{>0}$, we introduce a predicate $k\text{-mon} : \mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$ that expresses k-monotonicity: [modified from De Cooman et al. 2005*b*]

$$\begin{aligned} k\text{-mon}\,\underline{P} \Leftrightarrow &\forall A :\subseteq \Omega ; \\ &\forall \mathscr{A} :\subseteq \wp\Omega \wedge 0 < |\mathscr{A}| \leq k ; \\ &\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}} (-1)^{|\mathscr{B}|} \cdot \underline{P}(A \cap \bigcap \mathscr{B}) \geq 0, \end{aligned} \tag{2.10}$$

where $\bigcap \emptyset := \Omega$. The definition immediately allows us to see that every k-monotone lower probability is ℓ-monotone for all $\ell : 1..k$. We call the A in this definition a generating event.

Interesting subtypes are (i) monotone lower probabilities ($k := 1$, cf. $(1.33)_{42}$), (ii) 2-monotone lower probabilities (e.g., those generated by lower and upper probability functions defined on Ω (de Campos et al. 1994)), and (iii) belief functions (completely monotone, i.e., k-monotone for all $k : \mathbb{N}_{>0}$).

Note that k-monotonicity implies neither normedness nor nonnegativity, properties that are almost always required, however. The k-monotonicity of all constant lower probabilities $\underline{P} : \mathscr{P}\mathscr{I}_\Omega \wedge \underline{P} = \alpha$, where α is some real number, proves this: For any nonempty set of events $\mathscr{A}$ from Ω, we see that

$$\begin{aligned} \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}} (-1)^{|\mathscr{B}|} \cdot \alpha &\propto \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}} (-1)^{|\mathscr{B}|} \\ &= \textstyle\sum_{\ell:0..|\mathscr{A}|} \sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge|\mathscr{B}|=\ell} (-1)^{|\mathscr{B}|} \\ &= \textstyle\sum_{\ell:0..|\mathscr{A}|} \left(\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge|\mathscr{B}|=\ell} 1\right) \cdot (-1)^{\ell} \\ &= \textstyle\sum_{\ell:0..|\mathscr{A}|} \binom{|\mathscr{A}|}{\ell} \cdot (-1)^{\ell} \cdot 1^{|\mathscr{A}|-\ell} \\ &= 0, \end{aligned}$$

The binomial theorem: for all $n : \mathbb{N}$ and $x, y : \mathbb{R}^2$, it holds that $(x+y)^n = \sum_{\ell:0..n} \binom{n}{\ell} \cdot x^{\ell} \cdot y^{n-\ell}$, with $\binom{n}{\ell} := \frac{n!}{(n-\ell)!\cdot\ell!}$.

due to the binomial theorem. Adding just normedness is enough to also let nonnegativity be satisfied; furthermore, k-monotonicity is then sufficient for coherence if $k \geq 2$ [De Cooman et al. 2005*b*].

Even though $(2.10)_\cap$ gives a feasible set of linear constraints, it is possible to greatly reduce the number of redundant constraints, leading to a more efficient program. We finish this subsection by carrying out this reduction.

First notice that every constraint in $(2.10)_\cap$ depends only on subsets of its generating event A:

$$\begin{aligned}\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}}(-1)^{|\mathscr{B}|}\cdot\underline{P}(A\cap\bigcap\mathscr{B}) &= \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}}(-1)^{|\mathscr{B}|}\cdot\underline{P}\big(A\cap\bigcap_{B:\mathscr{B}}(A\cap B)\big)\\ &= \textstyle\sum_{\mathscr{B}:\subseteq\{A\cap B|B:\mathscr{A}\}}(-1)^{|\mathscr{B}|}\cdot\underline{P}(A\cap\bigcap\mathscr{B})\end{aligned}$$

This derivation shows that for each generating set A of Ω, we need only look at sets $\mathscr{A}:\subseteq\wp A$.

Also notice that whenever $A\in\mathscr{A}$, the left-hand sum in $(2.10)_\cap$ becomes identically zero:

$$\begin{aligned}&\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}}(-1)^{|\mathscr{B}|}\cdot\underline{P}(A\cap\bigcap\mathscr{B})\\ &\quad=\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge A\notin\mathscr{B}}(-1)^{|\mathscr{B}|}\cdot\underline{P}(A\cap\bigcap\mathscr{B})+\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge A\in\mathscr{B}}(-1)^{|\mathscr{B}|}\cdot\underline{P}(A\cap\bigcap\mathscr{B})\\ &\quad=\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge A\notin\mathscr{B}}(-1)^{|\mathscr{B}|}\cdot\big(\underline{P}(A\cap\bigcap\mathscr{B})-\underline{P}(A\cap A\cap\bigcap\mathscr{B})\big)\\ &\quad=0.\end{aligned}$$

Therefore, the constraint is trivially satisfied in this case, which means that we only need to look at sets $\mathscr{A}:\subseteq(\wp A)_{\neq A}$. This in turn implies that we do not need to consider the case $A:=\emptyset$.

Let us write out the intermediate expression for the definition of k-monotonicity we have now obtained:

$$\begin{aligned}k\text{-mon}\,\underline{P}\Leftrightarrow\forall A:(\wp\Omega)_{\neq\emptyset};&\\ \forall\mathscr{A}:\subseteq(\wp A)_{\neq A}\wedge 0<|\mathscr{A}|\leq k;&\\ \underline{P}A+\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge\mathscr{B}\neq\emptyset}(-1)^{|\mathscr{B}|}\cdot\underline{P}(\bigcap\mathscr{B})\geq 0.&\end{aligned}\tag{2.11}$$

The term $\underline{P}A$ we have written separately corresponds to $\mathscr{B}:=\emptyset$.

Suppose we separately consider 1-monotonicity, or monotonicity for short. It is expressed by a predicate $\text{mon}:\mathscr{P}\mathscr{I}_\Omega\to\mathbb{B}$ with a definition derived from the intermediate expression above:

$$\begin{aligned}\text{mon}\,\underline{P}\Leftrightarrow\forall A:(\wp\Omega)_{\neq\emptyset};&\\ \forall B:\subset A;&\\ \underline{P}A\geq\underline{P}B.&\end{aligned}\tag{2.12}$$

Then, for $\mathscr{A}:\subseteq\wp A$ such that $\bigcup\mathscr{A}\neq A$, the corresponding constraint in (2.11) is implied by monotonicity – i.e., $\underline{P}A\geq\underline{P}(\bigcup\mathscr{A})$ – and the constraint for $\bigcup\mathscr{A}$ as the generating event, because $\mathscr{A}\subseteq\wp(\bigcup\mathscr{A})$. So, this corresponding constraint may be dropped; or, in other words, we only need to consider sets $\mathscr{A}:\subseteq\wp A$ such that $\bigcup\mathscr{A}=A$, which are nonempty by definition. The same transitivity argument allows us to reduce the

number of constraints that have to be checked for monotonicity itself to the ones for which $|B| = |A| - 1$:

$$\begin{aligned} \operatorname{mon} \underline{P} \Leftrightarrow \forall A : (\wp\Omega)_{\neq\emptyset};\\ \forall B :\subset A \wedge |B| = |A| - 1;\\ \underline{P}A \geq \underline{P}B. \end{aligned} \tag{2.13}$$

Here, you can already see the typical exponential increase in the number of constraints with $|\Omega|$ which we alluded to in the next-to-last paragraph before §2.2$_{70}$.

The last set of redundant constraints we consider in the definition of k-monotonicity is related to the case when there is some $C : \mathscr{A}$ such that $C = \bigcap \mathscr{A}$. We can then write

$$\begin{aligned} &\textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge\mathscr{B}\neq\emptyset} (-1)^{|\mathscr{B}|} \cdot \underline{P}(\bigcap\mathscr{B})\\ &\quad = \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}_{\neq C}\wedge\mathscr{B}\neq\emptyset} (-1)^{|\mathscr{B}|} \cdot \underline{P}(\bigcap\mathscr{B}) + \sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge C\in\mathscr{B}} (-1)^{|\mathscr{B}|} \cdot \underline{P}C\\ &\quad = \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}_{\neq C}\wedge\mathscr{B}\neq\emptyset} (-1)^{|\mathscr{B}|} \cdot \underline{P}(\bigcap\mathscr{B}) - \underline{P}C \cdot \sum_{\mathscr{B}:\subseteq\mathscr{A}_{\neq C}} (-1)^{|\mathscr{B}|}\\ &\quad = \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}_{\neq C}\wedge\mathscr{B}\neq\emptyset} (-1)^{|\mathscr{B}|} \cdot \underline{P}(\bigcap\mathscr{B}), \end{aligned}$$

where the binomial theorem was used again in the last step. So, we only need to consider sets $\mathscr{A} :\subseteq \wp A$ such that $\bigcap\mathscr{A} \notin \mathscr{A}$. The special case of $\mathscr{A} :\subseteq \wp A$ that contain the empty set $\emptyset$ need therefore also not be considered.

We have finally arrived at a definition of k-monotonicity that leads to a more efficient program by eliminating a large number of redundant constraints in advance:

$$\begin{aligned} k\text{-}\operatorname{mon}\underline{P} \Leftrightarrow \operatorname{mon}\underline{P}\\ \wedge \forall A : (\wp\Omega)_{\neq\emptyset};\\ \forall \mathscr{A} :\subseteq \wp A \setminus \{\emptyset, A\} \wedge 2 \leq |\mathscr{A}| \leq k \wedge \begin{cases} \bigcup\mathscr{A} = A \\ \bigcap\mathscr{A} \notin \mathscr{A}; \end{cases}\\ \underline{P}A + \textstyle\sum_{\mathscr{B}:\subseteq\mathscr{A}\wedge\mathscr{B}\neq\emptyset} (-1)^{|\mathscr{B}|} \cdot \underline{P}(\bigcap\mathscr{B}) \geq 0. \end{aligned} \tag{2.14}$$

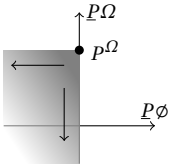
Constraints for $\Omega = \{a, b, c\}$, $k = 2$:
$\underline{P}\Omega \overset{*}{\geq} \underline{P}\{a, b\} \overset{*}{\geq} \underline{P}a \geq \underline{P}\emptyset$,
$\underline{P}\Omega \overset{*}{\geq} \underline{P}\{a, b\} + \underline{P}c - \underline{P}\emptyset$,
$\underline{P}\Omega \geq \underline{P}\{a, b\} + \underline{P}\{a, c\} - \underline{P}a$,
$\underline{P}\{a, b\} \geq \underline{P}a + \underline{P}b - \underline{P}\emptyset$.
(up to permutation; $\overset{*}{\geq}$ indicates a redundant constraint)

To get a feeling for the computational complexity of this definition, let us consider the number of candidate sets of events $\mathscr{A}$ when $A := \Omega$ and $k := |\Omega| - 1$: there are a total of $2^{2^{|\Omega|}-2}$ possible $\mathscr{A}$, which reduces to $\sum_{\ell:2..|\Omega|-1} \binom{2^{|\Omega|}-2}{\ell}$ when considering the constraints on $|\mathscr{A}|$. Let us illustrate this with some magnitudes for both quantities:
$|\Omega| = 4$: 10^4 and 10^3 (easy for current computers),
$|\Omega| = 5$: 10^9 and 10^4,
$|\Omega| = 6$: 10^{19} and 10^7 (hard, even with an efficient implementation).
An important difference with the original definition (2.10)$_{73}$ is that these numbers now go down sharply with $|A|$, whereas in the original definition they are valid for all $A :\subseteq \Omega$.

2.2.4 *Avoiding sure loss*

In this section's introduction, we have already drawn attention to the fact that the set of constraints for avoiding sure loss defined by $(2.2)_{67}$ is infinite and thus infeasible. In the current subsection, we first give a quickly obtained direct, vertex-based definition. After that, we still make the effort of finding a feasible constraint-based definition (i.e., one that generates a finite number of constraints); this allows us to combine avoiding sure loss with other properties and prepares us for work that cannot be avoided when looking at coherence in the next subsection.

The set $(\mathscr{P}\mathscr{I}_\Omega)_{\text{asl}}$ for $|\Omega| = 1$ (shaded) with its one extreme point and two extreme rays:



Walley [1991, §3.3.3$_{134}$] has shown that a lower prevision avoids sure loss if and only if it is dominated by some linear probability. This implies that the set of extreme sure-loss-avoiding lower probabilities consists of the degenerate probabilities (the extreme linear probabilities, cf. $(1.39)_{43}$) as extreme points and all negative main directions in $\wp\Omega \to \mathbb{R}$ as extreme rays. An alternative to definition $(1.25)_{40}$ for avoiding sure loss therefore becomes (let $\underline{P} : \mathscr{P}\mathscr{I}_\Omega$)

$$\begin{aligned} \text{asl}\,\underline{P} \Leftrightarrow & \exists \lambda : \Delta_\Omega; \\ & \exists \mu : (\mathbb{R}_{\leq 0})^{\wp\Omega}; \\ & \qquad \underline{P} = \textstyle\sum_{\omega:\Omega} \lambda_\omega \cdot P^\omega + \sum_{B:\subseteq\Omega} \mu_B \cdot I^{\imath B}. \end{aligned} \tag{2.15}$$

To find a feasible constraint-based definition, we start from the expression $(2.1)_{67}$ we used when introducing the concept of constraints in §2.1.1$_{67}$; we immediately replace indicators by events, write out the marginal gambles, and take into account the finite nature of Ω:

$$\begin{aligned} \text{asl}\,\underline{P} \Leftrightarrow & \forall \mathscr{B} :\subseteq \wp\Omega; \\ & \forall \lambda : (\mathbb{R}_{>0})^{\mathscr{B}}; \\ & \qquad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B \leq \sup\{\sum_{B:\mathscr{B}} \lambda_B \cdot I^B\}. \end{aligned} \tag{2.16}$$

Walley [1991, §A2$_{597}$] does remark that the nonnegativity assumption is not substantive.

Without giving an elaborate proof and with the additional assumption of nonnegativity – i.e., that $\underline{P}$ is an element of the first orthant of $\wp\Omega \to \mathbb{R}$ –, Walley [1991, §A3$_{598}$] mentions how the corresponding set of constraints can be reduced to a finite one. We do not assume nonnegativity.

First of all, the cases $\mathscr{B} := \imath\emptyset$ and $\mathscr{B} := \imath\Omega$ lead to the simple constraints $\underline{P}\emptyset \leq 0$ and $\underline{P}\Omega \leq 1$, which we consider separately from the rest. The reason is that for any $\underline{P}$ that satisfies these constraints, all other constraints involving $\emptyset$ or Ω are made redundant by the corresponding (equivalent or more stringent) constraint not involving $\emptyset$ or Ω. Looking at the inequality in (2.16), this follows for $\emptyset$ from the fact that the left-hand side cannot increase ($\underline{P}\emptyset \leq 0$) and the right-hand side stays constant ($I^\emptyset = 0$). For Ω this follows from the fact that the left-hand side decreases less ($\underline{P}\Omega \leq 1$) than the right-hand side ($I^\Omega = 1$).

Each of the other constraints can be replaced by an equivalent or

more stringent constraint written in some standard form that we now set out to identify.

The first thing we do is rescale the coefficients in such a way that the right-hand side supremum of the constraint in (2.16) is 1. So we replace λ by $\lambda / \sup\{\sum_{B:\mathscr{B}} \lambda_B \cdot I^B\}$, which has components in $]0,1]$. This creates an equivalent constraint. Assume from now on that this rescaling has been done for all λ.

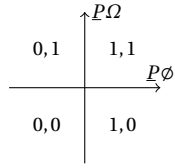
This rescaling is only the first step on our road to the standard form. One of the other things we are going to do is fiddle with the linear combinations appearing in the left-hand and right-hand sides of the constraint in (2.16): We will increase the coefficients λ and enlarge the set the sum ranges over, but we do this in such a way as to keep the right-hand side unchanged. This will result in constraints that are still part of the set described by (2.16), but we need to ensure that it results in constraints that are either equivalent to or more stringent than the ones we started out with. In that case, we can drop the original constraints, as they are implied by the new ones we replace them with.

For lower probabilities $\underline{P}$ that are nonnegative everywhere, the left-hand side increases when fiddling (as we have described just above) with the linear combination, whereas the right-hand side remains constant (equal to 1), so the constraint becomes more stringent. However, when $\underline{P}$ has negative components, the left-hand side might decrease, causing a possibly less stringent constraint. To counter this, we use binary masks $\beta : \mathbb{B}^{\wp\Omega}$ to create a specific constraint for all possible orthants $\underline{P}$ can be located in. This, together with the rescaling and the separate treatment of $\iota\emptyset$ and $\iota\Omega$, but without the fiddling, results in the following intermediate equivalent expression for the definition of avoiding sure loss:

$$\begin{aligned}
\operatorname{asl}\underline{P} \Leftrightarrow\ & \underline{P}\emptyset \leq 0 \wedge \underline{P}\Omega \leq 1 \\
& \wedge \forall \beta : \mathbb{B}^{\wp\Omega}; \\
& \quad \forall \mathscr{B} :\subseteq \wp\Omega \setminus \{\emptyset, \Omega\}; \\
& \quad \forall \lambda :]0,1]^{\mathscr{B}} \wedge \sup\{\textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B\} = 1; \\
& \qquad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \beta_B \cdot \underline{P}B \leq 1.
\end{aligned} \tag{2.17}$$

Let us explain how these binary masks work: Whenever $\underline{P}B < 0$ for some $B : \mathscr{B}$, there is a constraint for which $\beta_B = 0$, which will be more stringent than the original constraint. We need to be cautious, however, because only when $\beta = 1$ (suitable for $\underline{P}$ in the first orthant, with only nonnegative components), we have a constraint that is of the form given by the inequality in (2.16). The others might be *too* stringent.

Binary masks β for $|\Omega| = 1$:



Luckily, however, these other constraints are never too stringent: We can show that the constraints for which $\beta \neq 1$ are also implied by the original definition. This follows by replacing λ in the original definition

by $\beta \cdot \lambda$ (the set of constraints implied by definition $(2.16)_{76}$ does not change when allowing coefficients with components equal to 0) and realizing that $\sup\{\sum_{B:\hat{\mathscr{B}}} \lambda_B \cdot \beta_B \cdot I^B\} \leq \sup\{\sum_{B:\hat{\mathscr{B}}} \lambda_B \cdot I^B\}$.

The ideas used here are already present in Walley [1991, §A2$_{597}$].

Now we start fiddling fearlessly with the linear combination: we only focus on the sum under the supremum appearing in $(2.17)_{\frown}$, as we have already taken care of the effect of the fiddling on (the other parts of) the constraint. We can 'fill things up', i.e., modify the sum in such a way that it results in the constant function $\sum_{B:\hat{\mathscr{B}}} \hat{\lambda}_B \cdot I^B = I^\Omega = 1$. In this modified sum, $\hat{\mathscr{B}} :\subseteq \wp\Omega$ is such that $\mathscr{B} \subseteq \hat{\mathscr{B}}$ and the coefficients $\hat{\lambda}$ in $(\mathbb{R}_{>0})^{\hat{\mathscr{B}}}$ are such that the rescaled $\lambda \leq \hat{\lambda}_{\mathscr{B}}$. This modification is done as follows:

(i) Let $\hat{\mathscr{B}} := \mathscr{B}$; we go over all B in $\hat{\mathscr{B}}$ and – for the first set B – increase λ_B to
$$\hat{\lambda}_B = \lambda_B + \min\{(1 - \textstyle\sum_{C:\hat{\mathscr{B}}} \lambda_C \cdot I^C)_B\};$$
for subsequent sets the increase is calculated similarly, but the appearing sum uses the increased coefficients where available.

(ii) If the procedure above does not yield a sum $\sum_{B:\hat{\mathscr{B}}} \hat{\lambda}_B \cdot I^B$ that is identically one we go over all B in $\wp\Omega \setminus \mathscr{B}$ and – for the first set B – add B to $\hat{\mathscr{B}}$ and set
$$\hat{\lambda}_B = \min\{(1 - \textstyle\sum_{C:\hat{\mathscr{B}}\setminus B} \hat{\lambda}_C \cdot I^C)_B\};$$
for subsequent sets the coefficient is again calculated similarly. Because $\wp\Omega$ is finite, this procedure is guaranteed to stop after a finite number of steps; at that point, the sum is identically one and $\hat{\mathscr{B}}$ only contains sets with coefficients that are nonzero.

The filling-up procedure is illustrated for the case $\Omega = \{a, b, c\}$ in the following example:

$\mathscr{B} := \{\{a,b\}, \imath b, \imath c\}$; $\lambda_{\{a,b\}} := \frac{1}{2}$, $\lambda_{\imath b} := \frac{1}{2}$, $\lambda_{\imath c} := \frac{2}{3}$; $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B =$ 

$\hat{\mathscr{B}} := \{\{a,b\}, \imath b, \imath c, \imath a\}$; $\hat{\lambda}_{\{a,b\}} := \frac{1}{2}$, $\hat{\lambda}_{\imath b} := \frac{1}{2}$, $\hat{\lambda}_{\imath c} := 1$, $\hat{\lambda}_{\imath a} := \frac{1}{2}$; $\sum_{B:\hat{\mathscr{B}}} \hat{\lambda}_B \cdot I^B =$ 

(2.18)

This procedure causes $\bigcup \hat{\mathscr{B}} = \Omega$ as a side-effect and also implies that the set of functions $\imath I^\Omega \cup \{I^B \mid B : \hat{\mathscr{B}}\}$ is linearly dependent.

A predicate $\mathrm{dep} :\wp\mathscr{L}_\Omega \to \mathbb{B}$ that expresses linear dependence of some finite set of gambles $\mathscr{K} :\subseteq \mathscr{L}_\Omega$ is therefore useful to have at our disposal. It is defined by

$$\begin{aligned} \mathrm{dep}\,\mathscr{K} \Leftrightarrow \exists \mu : \mathbb{R}^{\mathscr{K}} \wedge \mu \neq \mathscr{K}\,;0; \\ \textstyle\sum_{f:\mathscr{K}} \mu_f \cdot f = 0. \end{aligned} \tag{2.19}$$

When talking about linear independence, we can use the (implicitly pointwise extended) negation $\neg\,\mathrm{dep}$ of this predicate.

Let us write out the intermediate expression for the definition of avoiding sure loss we now obtain (dropping hats):

$$\begin{aligned}
&\operatorname{asl}\underline{P} \Leftrightarrow \underline{P}\varnothing \leq 0 \wedge \underline{P}\Omega \leq 1 \\
&\quad \wedge \forall \beta : \mathbb{B}^{\wp\Omega}; \\
&\qquad \forall \mathscr{B} :\subseteq \wp\Omega \setminus \{\varnothing, \Omega\} \mathbin{\text{a}\wedge} \bigcup \mathscr{B} = \Omega \wedge \operatorname{dep}\big(\imath I^{\Omega} \cup \{I^B \mid B : \mathscr{B}\}\big); \qquad (2.20) \\
&\qquad \forall \lambda :]0,1]^{\mathscr{B}} \mathbin{\text{a}\wedge} \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = 1; \\
&\qquad\quad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \beta_B \cdot \underline{P}B \leq 1.
\end{aligned}$$

This definition is still infeasible; for general $\mathscr{B}$, the set of possible coefficients λ is still infinite. In the filling-up example (2.18), the coefficients $\hat{\lambda}_{\{a,b\}}$, $\hat{\lambda}_{\imath a}$, and $\hat{\lambda}_{\imath b}$ were chosen equal to 1/2, but they only needed to satisfy $\hat{\lambda}_{\imath a} = \hat{\lambda}_{\imath b} = 1 - \hat{\lambda}_{\{a,b\}}$.

Technical lemma (A.1)–(A.2)$_{198}$ – applied with $q := \beta \cdot \underline{P}$ and $\rho := 1$ – allows us to derive that we only need to consider those sets of events $\mathscr{B} :\subseteq \wp\Omega$ such that the set of gambles $\{I^B - \beta_B \cdot \underline{P}B \mid B : \mathscr{B}\}$ is linearly independent and those coefficients $\lambda : \mathbb{R}_{>0}{}^{\mathscr{B}}$ for which $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B$ is constant. Because $\lambda > 0$, we can implicitly rescale this constant to 1, which gives us back the requirement $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = 1$. This in turn allows us to use technical lemma (A.3)$_{199}$ to show that linear independence of $\{I^B - \beta_B \cdot \underline{P}B \mid B : \mathscr{B}\}$ is equivalent to the linear independence of $\{I^B \mid B : \mathscr{B}\}$.

Linear dependence of $\imath I^{\Omega} \cup \{I^B \mid B : \mathscr{B}\}$ ensures that the linear system of equations $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = 1$ has a solution λ. Linear independence of $\{I^B \mid B : \mathscr{B}\}$ makes sure that the number of unknowns $|\mathscr{B}|$ is never larger than the number of equations $|\Omega|$, so this solution is unique. As the coefficients in the system are $\mathbb{B}$-valued, the unique solution is rational. This is what allows us to formulate a feasible constraint-based definition, as there are only a finite number of possible $\mathscr{B}$, and we must only consider one rational-valued λ per $\mathscr{B}$.

So we have finally found a feasible definition for avoiding sure loss:

$$\begin{aligned}
&\operatorname{asl}\underline{P} \Leftrightarrow \underline{P}\varnothing \leq 0 \wedge \underline{P}\Omega \leq 1 \\
&\quad \wedge \forall \mathscr{B} :\subseteq \wp\Omega \setminus \{\varnothing, \Omega\} \mathbin{\text{a}\wedge} \bigcup \mathscr{B} = \Omega \wedge \begin{cases} \operatorname{dep}\big(\imath I^{\Omega} \cup \{I^B \mid B : \mathscr{B}\}\big) \\ \neg \operatorname{dep}\{I^B \mid B : \mathscr{B}\}; \end{cases} \\
&\qquad \forall \lambda : (\mathbb{Q} \cap]0,1])^{\mathscr{B}} \mathbin{\text{a}\wedge} \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = 1; \qquad (2.21) \\
&\qquad \forall \beta : \mathbb{B}^{\mathscr{B}}; \\
&\qquad\quad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \beta_B \cdot \underline{P}B \leq 1.
\end{aligned}$$

Let $\Omega := \{a, b, c\}$. The linear system of equations for $\mathscr{B} := \big\{\{a, b\}, \{a, c\}, \{b, c\}\big\}$,

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_{\{a,b\}} \\ \lambda_{\{a,c\}} \\ \lambda_{\{b,c\}} \end{pmatrix} = 1,$$

where $\lambda : (\mathbb{R}_{>0})^{\mathscr{B}}$, results in $\lambda = 1/2$.

Constraints for $\Omega = \{a, b, c\}$:
$\underline{P}\varnothing \leq 0$, $\underline{P}\Omega \leq 1$,
$\underline{P}a + \underline{P}b + \underline{P}c \leq 1$,
$\underline{P}\{a, b\} + \underline{P}c \leq 1$.
$\underline{P}\{a, b\} + \underline{P}\{a, c\} + \underline{P}\{b, c\} \overset{*}{\leq} 2$,
(unmasked and up to permutation; $\overset{*}{\leq}$ indicates a redundant constraint)

Note that the quantification over λ just selects the unique solution. The binary masks are applied at the end out of efficiency considerations; checking linear (in)dependence and solving for λ are the most computationally intensive steps of the constraint generation process. The large

number of redundant constraints the binary masks create in practice are easily removed by redundant constraint removal algorithms.

2.2.5 *Coherence*

We have stated in §1.2.5$_{41}$ that coherence is the strongest universal requirement we want to impose on lower previsions, and thus on lower probabilities. Therefore, a feasible constraint-based definition for the coherence of lower probabilities is the most important of all the definitions in this section. Again, Walley [1991, §A3$_{598}$] mentions a finitary constraint-based definition without elaborating.

Our starting point is the original definition (1.28)$_{41}$, which we immediately reformulate for lower probabilities $\underline{P} : \mathscr{P}\mathscr{I}_\Omega$ on a finite possibility space Ω; i.e., we work with events instead of gambles. For this, we have to write out the marginal gambles using (1.20)$_{37}$; we obtain

$$\begin{aligned}
&\operatorname{coh}\underline{P} \Leftrightarrow \operatorname{asl}\underline{P}\\
&\qquad \wedge \forall A :\subseteq \Omega;\\
&\qquad\quad \forall \mathscr{B} :\subseteq (\wp\Omega)_{\neq A};\\
&\qquad\quad \forall \lambda : (\mathbb{R}_{>0})^{\mathscr{B}\cup\imath A};\\
&\qquad\qquad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B - \lambda_A \cdot \underline{P}A \leq \sup\{\sum_{B:\mathscr{B}} \lambda_B \cdot I^B - \lambda_A \cdot I^A\},
\end{aligned} \tag{2.22}$$

where we have also added a coefficient corresponding to A; it does not change the definition, which can be seen by dividing both sides of the inequality by λ_A.

The form of each of the (infinite number of) constraints for coherence is very similar to the one for avoiding sure loss (2.16)$_{76}$, but there are differences. On the one hand, those differences imply that coherent probabilities are nonnegative, so the trick with the binary masks is not necessary; on the other hand, they give rise to a number of different cases we need to look at separately. Once each of these cases is formulated, the ideas we used for reformulating the definition of avoiding sure loss can be straightforwardly applied.

First of all, let us look at the two simple cases:

(i) Applying avoiding sure loss (1.26)$_{40}$ to the gambles in $\mathscr{I}_\Omega$, we find $\underline{P}B \leq \sup\{I^B\}$, so $\underline{P}B \leq 1$ for all nonempty B of Ω and $\underline{P}\emptyset \leq 0$.

(ii) Taking $\mathscr{B} = \emptyset$, we find the constraint $\underline{P}A \geq \inf\{I^A\}$, so $\underline{P}A \geq 0$ for all $A :\subset \Omega$ and $\underline{P}\Omega \geq 1$.

This means that a coherent lower probability $\underline{P}$ must be $[0,1]$-bounded and normed, which implies nonnegativity. Normedness also implies we do not have to consider $\emptyset$ and Ω any more as candidates for A or elements of $\mathscr{B}$; they result in constraints that are equivalent to the corresponding constraints not involving $\emptyset$ or Ω.

We now show how to write all constraints in one of three standard forms, each of which will be investigated separately afterwards.

The right-hand side supremum is either positive, zero, or negative. So, by normalizing λ and filling things up, we can respectively make the right-hand side sum, and thus supremum, identically 1, 0, or -1. This time, the possibilities for filling up not only include the ones we encountered when looking at avoiding sure loss in the previous subsection, to wit, (take $\hat{\mathscr{B}} :\subseteq \wp\Omega \wedge \mathscr{B} \subseteq \hat{\mathscr{B}}$ and $\hat{\lambda} : (\mathbb{R}_{>0})^{\hat{\mathscr{B}} \cup \imath A} \wedge \lambda_{\mathscr{B}} \leq \hat{\lambda}_{\mathscr{B}} \wedge \lambda_A \geq \hat{\lambda}_A$)

(i) appropriately increasing the coefficients to $\hat{\lambda}_B$ for some well-chosen events $B : \mathscr{B}$,

(ii) adding some well-chosen events $C : \wp\Omega \setminus \mathscr{B}$ – creating the set of events $\hat{\mathscr{B}}$ – and appropriately setting the value of the corresponding coefficients $\hat{\lambda}_C$,

but they also include

(iii) appropriately decreasing the coefficient λ_A to $\hat{\lambda}_A$.

Whenever this decrease would lead to a nonpositive $\hat{\lambda}_A$, we effectively get a constraint for avoiding sure loss, something we are already taking into account.

The effect of this procedure on the constraint as a whole is that, due to the nonnegativity of $\underline{P}$, the left-hand side can only increase while the right-hand side stays constant (1, 0, or -1): A more stringent constraint is obtained. We therefore only need to consider the constraints resulting from this procedure (dropping hats):

$$\begin{aligned} \operatorname{coh} \underline{P} \Leftrightarrow \operatorname{asl} \underline{P} \wedge \operatorname{nrm} \underline{P} \wedge 0 \leq \underline{P} \leq 1 \\ \wedge \forall A : \wp\Omega \setminus \{\emptyset, \Omega\}; \\ \forall \mathscr{B} :\subseteq \wp\Omega \setminus \{\emptyset, A, \Omega\}; \\ \forall \gamma : \{1, 0, -1\}; \\ \forall \lambda : (\mathbb{R}_{>0})^{\mathscr{B} \cup \imath A} \wedge \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B - \lambda_A \cdot I^A = \gamma; \\ \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B - \lambda_A \cdot \underline{P}A \leq \gamma. \end{aligned} \tag{2.23}$$

We first consider constraints with $\gamma = 1$. This situation can only occur when at the same time $\{I^\Omega, I^A\} \cup \{I^B \mid B : \mathscr{B}\}$ is a linearly dependent set and $\bigcup \mathscr{B} = \Omega$. Lemma (A.1)–(A.2)$_{198}$ applied with $q := \underline{P}$ and $\rho := 1$ allows us to derive that we only need to consider those A and $\mathscr{B}$ such that the set of gambles $\imath(I^A - \underline{P}A) \cup \{I^B - \underline{P}B \mid B : \mathscr{B}\}$ is linearly independent and those coefficients $\lambda : \mathbb{R}_{>0}{}^{\mathscr{B} \cup \imath A}$ for which $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B - \lambda_A \cdot I^A$ is constant. After rescaling, this constant can be 1, 0, or -1. Generated constraints for the last two cases are treated further on; for the first case, we can then use lemma (A.3)$_{199}$ to show that linear independence of the given set is equivalent to linear independence of $\imath I^A \cup \{I^B \mid B : \mathscr{B}\}$. This ensures there is a unique (rational-valued) solution λ to the given condition.

Next, we treat constraints with $\gamma = 0$. This situation can only occur when $\bigcup\mathscr{B} = A$. We can normalize λ such that $\lambda_A = 1$ and such that the other components lie in $]0,1]$, satisfying $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = I^A$, which in turn implies $\imath I^A \cup \{I^B \mid B:\mathscr{B}\}$ is linearly dependent. The form of the constraint becomes $\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B \leq \underline{P}A$. Note that this form can be seen as avoiding sure loss relative to A (cf. $(2.16)_{76}$); in our final definition, we shall therefore merge avoiding sure loss into this case by allowing A to be Ω. Similarly as when deriving the feasible definition of avoiding sure loss, we can use lemma $(A.1)$–$(A.2)_{198}$, now relative to A and applied with $q := \underline{P}_{\wp A}$, to derive that we only need to consider those $\mathscr{B}$ for which $\{I^B - \underline{P}B \mid B:\mathscr{B}\}$ is linearly independent. Because of the condition $\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = I^A$ on λ, lemma $(A.3)_{199}$ – again applied relative to A – shows that this is already guaranteed when $\{I^B \mid B:\mathscr{B}\}$ is linearly independent. This again ensures there is a unique (rational-valued) solution λ to the given condition.

Finally, we consider constraints with $\gamma = -1$. They can only occur when $A = \Omega$, which we have already excluded, so this case can be ignored.

Incorporating what we have learned about the three different cases results in a feasible definition for the coherence of a lower probability:

Constraints for $\Omega = \{a,b,c\}$:
$\underline{P}\emptyset = 0, \underline{P}\Omega = 1,$
$0 \leq \underline{P}a \overset{*}{\leq} 1, 0 \overset{*}{\leq} \underline{P}\{a,b\} \overset{*}{\leq} 1$
$\underline{P}\{a,b\} + \underline{P}\{a,c\} - \underline{P}a \leq 1.$
$\underline{P}a + \underline{P}b + \underline{P}c \overset{*}{\leq} \underline{P}\Omega,$
$\underline{P}\{a,b\} + \underline{P}c \overset{*}{\leq} \underline{P}\Omega.$
$\underline{P}\{a,b\} + \underline{P}\{a,c\} + \underline{P}\{b,c\} \overset{*}{\leq} 2 \cdot \underline{P}\Omega,$
$\underline{P}a + \underline{P}b \leq \underline{P}\{a,b\},$
(up to permutation; $\overset{*}{\leq}$ indicates a redundant constraint)

$$
\begin{aligned}
&\operatorname{coh}\underline{P} \Leftrightarrow \operatorname{nrm}\underline{P} \\
&\qquad \wedge 0 \leq \underline{P} \leq 1 \\
&\qquad \wedge \forall A:\wp\Omega \setminus \{\emptyset,\Omega\}; \\
&\qquad\quad \forall \mathscr{B}:\subseteq \wp\Omega \setminus \{\emptyset, A, \Omega\} \wedge \bigcup\mathscr{B} = \Omega \wedge \begin{cases} \operatorname{dep}\big(\{I^\Omega, I^A\} \cup \{I^B \mid B:\mathscr{B}\}\big) \\ \neg \operatorname{dep}\big(\imath I^A \cup \{I^B \mid B:\mathscr{B}\}\big); \end{cases} \\
&\qquad\quad \forall \lambda : (\mathbb{Q}_{>0})^{\mathscr{B} \cup \imath A} \wedge \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B - \lambda_A \cdot I^A = 1; \\
&\qquad\qquad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B - \lambda_A \cdot \underline{P}A \leq 1 \\
&\qquad \wedge \forall A:\wp\Omega_{\neq\emptyset}; \\
&\qquad\quad \forall \mathscr{B}:\subseteq \wp A \setminus \{\emptyset, A\} \wedge \bigcup\mathscr{B} = A \wedge \begin{cases} \operatorname{dep}\big(\imath I^A \cup \{I^B \mid B:\mathscr{B}\}\big) \\ \neg \operatorname{dep}\{I^B \mid B:\mathscr{B}\}; \end{cases} \\
&\qquad\quad \forall \lambda : (\mathbb{Q} \cap]0,1])^{\mathscr{B}} \wedge \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot I^B = I^A; \\
&\qquad\qquad \textstyle\sum_{B:\mathscr{B}} \lambda_B \cdot \underline{P}B \leq \underline{P}A.
\end{aligned}
\tag{2.24}
$$

It is again interesting to get some feeling for the computational complexity of this definition. So let us consider the number of candidate sets of events $\mathscr{B}$ for the last case and for $A := \Omega$, i.e., for avoiding sure loss: there are a total of $2^{2^{|\Omega|}-2}$ possible $\mathscr{B}$, which reduces to $\sum_{\ell:2..|\Omega|} \binom{2^{|\Omega|}-2}{\ell}$ when considering the requirement $\bigcup\mathscr{B} = \Omega$ (when $\ell > 1$) and the linear independence requirement (when $\ell \leq |\Omega|$). Let us illustrate this with some magnitudes for both quantities:
$|\Omega| = 4$: 10^4 and 10^3 (easy for current computers),
$|\Omega| = 5$: 10^9 and 10^5,

$|\Omega| = 6$: 10^{18} and 10^8 (hard, even with an efficient implementation). Generating the sets corresponding to the second quantity, as well as their large number – as each implies a number of subsequent processing steps – are the main contributors to the computational complexity.

With definition (2.24), we have given the most important theoretical result in this chapter. This does not mean that all is said and done on the theoretical side: Before talking about the more practical results, we treat two other interesting topics in the next two subsections.

2.2.6 *Permutation invariance*

Up until now, all properties we have looked at were criteria that more or less defined the set of all possible lower probabilities one would consider working with in general: rationality criteria (avoiding sure loss, coherence) or criteria justifiable on the basis of their mathematical simplicity (additivity, k-monotonicity). There are other properties that can be used to restrict attention to a subset of lower probabilities that are appropriate in specific cases.

In this subsection, we look at the case where we (have reasons to) assume that the uncertainty model we use should be invariant under a set of permutations of the possibility space. This assumption typically arises in a situation where we have evidence that is symmetrical; e.g., about a six-faced die we get trustworthy information that every face comes up at least one tenth of the time. This does not preclude the die from being loaded, e.g., of one face coming up half of the time. So this *weak* permutation invariance differs from the *strong* permutation invariance which occurs when we have evidence of symmetry; e.g., when the die is tested and proclaimed fair.

De Cooman & Miranda [2007] discuss uncertainty models for symmetrical situations and symmetrical uncertainty models, which include the weakly permutation invariant previsions.

When considering the set of all permutations

$$\Pi_\Omega := \left\{ \pi : \Omega \cup \wp\Omega \leftrightarrow \Omega \cup \wp\Omega \;\middle|\; \begin{array}{l} \forall \omega : \Omega ; \pi\omega \in \Omega \\ \wedge \forall A :\subseteq \Omega ; \pi A = \{\pi\omega \mid \omega : A\} \end{array} \right\}. \tag{2.25}$$

of elements and subsets of a finite possibility space Ω, a lower probability $\underline{P}$ on $\wp\Omega$ is called (weakly) permutation invariant when the lower probability $\underline{P}A$ of an event A of Ω is the same as the lower probability for any permuted event. Permutation invariance is then formalized using the predicate $\mathrm{pin} : \mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$ defined by

$$\begin{aligned} \mathrm{pin}\,\underline{P} \Leftrightarrow \forall \pi : \Pi_\Omega ; \\ \forall A :\subseteq \Omega ; \\ \underline{P}(\pi A) = \underline{P}A. \end{aligned} \tag{2.26}$$

As any permutation preserves the cardinality of the event, and as we consider the set of all permutations, a weakly permutation invariant lower probability must be constant on events with constant cardinality.

Also (let $B, C :\subseteq \Omega^2$), because the constraints $\underline{P}B = \underline{P}A$ and $\underline{P}C = \underline{P}A$ plus transitivity of equality make the constraint $\underline{P}B = \underline{P}C$ redundant, we can reduce the number of nonredundant constraints by keeping one side of the equality fixed. This is done by ordering the events – any strict total order $\prec : (\wp\Omega)^2 \to \mathbb{B}$ will do – and letting A be the minimum of the relative order for each cardinality. We get [also see Weichselberger 2001, §4.3.1$_{488}$]

Example orderings of $\wp\{a,b,c\}$. Lexicographical: $\emptyset \prec \imath a \prec \{a,b\} \prec \{a,b,c\} \prec \{a,c\} \prec \imath b \prec \{b,c\} \prec \imath c$. Cardinality-then-lexicographical: $\emptyset \prec \imath a \prec \imath b \prec \imath c \prec \{a,b\} \prec \{a,c\} \prec \{b,c\} \prec \{a,b,c\}$.

$$\begin{aligned} &\operatorname{pin}\underline{P} \Leftrightarrow \forall k : 1..|\Omega| - 1; \\ &\qquad \forall B :\subset \Omega \wedge |B| = k \wedge B \succ \min\{A :\subset \Omega \mid k = |A|\}; \\ &\qquad\qquad \underline{P}B = \underline{P}(\min\{A :\subset \Omega \mid k = |A|\}). \end{aligned} \tag{2.27}$$

Programming-wise, using an ordering of the events is very natural, as most objects (such as the power set $\wp\Omega$) are encoded using list-like structures.

2.2.7 *Maxitivity*

The last property we are going to look at is the one that is characteristic for necessity measures, which are the conjugate lower probabilities of the (maxitive) possibility measures [Dubois & Prade 1988, §1.3$_{13}$]. These necessity measures are a special type of belief functions (cf. (2.8)$_{72}$), but we are going to make abstraction of that fact here.

An upper probability $\overline{P}$ on $\wp\Omega$ is maxitive when its value in the union of a couple of events B and C of Ω is the maximum of the values in each of these events ($\overline{P}(B \cup C) = \max\{\overline{P}B, \overline{P}C\}$), and therefore – for finite Ω – the maximum of the values in the union's elementary events. So an upper probability $\overline{P}$ is maxitive if

$$\begin{aligned} &\forall A :\subseteq \Omega \wedge |A| > 1; \\ &\qquad \overline{P}A = \max\{\overline{P}_A\}. \end{aligned} \tag{2.28}$$

Note the similarity to the additivity property (2.6)$_{71}$. To avoid technical issues, we adopt the ad hoc convention that $\overline{P}\emptyset \leq \overline{P}_\Omega$, i.e., the value in the empty set is not uniquely defined, but only constrained from above.

We use the conjugacy property (1.15)$_{37}$ to translate this maxitivity property to lower probabilities and obtain a predicate $\operatorname{nec} :\mathscr{P}\mathscr{I}_\Omega \to \mathbb{B}$ defined by

$$\begin{aligned} &\operatorname{nec}\underline{P} \Leftrightarrow \forall A :\subset \Omega \wedge |A| < |\Omega| - 1; \\ &\qquad \underline{P}A = \min_{\omega :\Omega \setminus A} \underline{P}\Omega_{\neq\omega}, \end{aligned} \tag{2.29}$$

now with the ad hoc convention $\underline{P}\Omega \geq \max_{\omega :\Omega} \underline{P}\Omega_{\neq\omega}$.

There is a problem with this property: it generates nonlinear constraints. Even worse, the predicate is nonconvex: take $\underline{P}' : \mathscr{P}\mathscr{I}_\Omega$, then $\operatorname{nec}\underline{P} \wedge \operatorname{nec}\underline{P}'$ does not guarantee $\operatorname{nec}\left(\frac{1}{2}\underline{P} + \frac{1}{2}\underline{P}'\right)$! This means the set of all lower probabilities satisfying nec is not convex.

So why *do* we investigate this property? A little reasoning allows us to still use our standard procedure of first generating linear constraints and then doing vertex enumeration to obtain the structure of the set of all lower probabilities satisfying nec. This is why we investgate this property.

Recall that in §2.2.4$_{76}$, for finding a feasible set of constraints for avoiding sure loss, we at some point used binary masks to introduce constraints for each orthant $\underline{P}$ could be located in. Here we do something similar: to be able to rewrite (2.29) using linear constraints, we have to be able to explicitly write out the appearing minima. These are fully determined when the set of values $\{\underline{P}\Omega_{\neq\omega} \mid \omega : \Omega\}$ can be totally ordered, which is always the case. We therefore consider all the possible orderings and transfer these to a strict total order of the events of cardinality $|\Omega| - 1$ to obtain

$$\begin{aligned} \operatorname{nec}\underline{P} \Leftrightarrow \exists \text{ strict total order } \prec : \{\Omega_{\neq\omega} \mid \omega : \Omega\}^2 \to \mathbb{B}; \\ \underline{P}\Omega \geq \underline{P}(\max\{\Omega_{\neq\omega} \mid \omega : \Omega\}) \\ \wedge \forall A :\subset \Omega \wedge |A| < |\Omega| - 1; \\ \underline{P}A = \underline{P}(\min\{\Omega_{\neq\omega} \mid \omega : \Omega \setminus A\}), \end{aligned} \tag{2.30}$$

where we have incorporated our ad hoc convention concerning the lower probability of the possibility space into the definition.

The trick we used to obtain (2.30) also tells us something about the structure of the set of lower probabilities satisfying nec: For every order, we obtain a specific convex set we call a lobe, so the whole set is a union of lobes. As the events of cardinality $|\Omega| - 1$ are determined by the missing elementary event, an ordering of these sets can therefore be characterized by an (inverse) ordering of the missing elementary events. This allows us to see that the set of all orders can be obtained by applying all permutations of Ω to any one order. This means that all lobes are identical up to a permutation of the possibility space Ω and we only need to compute one.

We have arrived at the end of our discussion of a number of interesting properties for lower probabilities and their constraint-based definition. Now it is time to leave the land of constraints and move to the land of extremes.

2.3 Results

In this chapter, we have started out by drawing attention to a nice property of linear probabilities: any one of them can be written as a convex combination of the degenerate probabilities, the extreme linear probabilities. We claimed that the same could be done for lower probabilities that are coherent or avoid sure loss; to wit, we claimed extreme lower proba-

bilities could be found for these cases. In §2.1$_{66}$, the proposed approach was shown: write down a constraint-based definition for the property of interest and then apply a vertex enumeration algorithm to find the extreme points corresponding to the generated set of constraints. In the previous section, we deduced feasible, sometimes relatively efficient constraint-based definitions for a number of interesting properties.

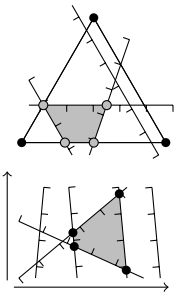
In this text's electronic version, the raw input and output files as well as the source of the 'constraints' program, are attached.

Here, in this section, we are going to look mainly at the output of the vertex enumeration step. Of course, you will not be served with the raw output of the programs we used [Avis 2000; Fukuda & Prudon 1996]; i.e., text file upon text file of ordered numbers. Nor will you be subjected to the raw input for these programs – text file upon text file of ordered numbers, again –, generated by the "constraints" program I wrote (encouraged by Matthias Troffaes's 2-monotonicity prototype). No, an illustrated overview of some simple results is far better to carry the important ideas across; more (extensive) results can be found in the Herbarium$_{194}$.

2.3.1 *Assorted general remarks*

Because we are still studying an aspect of *normative* probability theory, we only present the extreme lower probabilities for sets of properties that include or imply the property of avoiding sure loss.

Most of the properties we have discussed in the previous section do not imply nonnegativity (2.5)$_{71}$ and normedness (2.4)$_{71}$ by themselves; for example: lower probabilities that avoid sure loss can have negative components and k-monotonicity is a scale-independent and translation-independent property. However, in this section, we add these properties if necessary, which ensures that all extreme lower probabilities $\underline{P}$ on $\wp\Omega$ are located in the first orthant ($\underline{P} \geq 0$) and have trivial values in the improper events ($\underline{P}\emptyset = 0$ and $\underline{P}\Omega = 1$). Avoiding sure loss then guarantees (cf. (1.26)$_{40}$) that all components of all extreme lower probabilities are bounded, so we only encounter polytopes.



In case the set of properties implies coherence, it is possible to give a graphical representation of the corresponding extreme lower probabilities by using their credal set. Now that we are talking about credal sets, it is also a good moment to draw attention to the almost entirely parallel construction of extreme lower probabilities and the extreme points of a credal set: the space of interest in the latter case is the unit simplex Δ_Ω instead of $\wp\Omega \to \mathbb{R}$ and every component of the lower probability for which we want to find the credal set specifies a constraint. (Compare the figure at the end of §1.2.9$_{50}$ with the one at the beginning of §2.1.2$_{68}$, both reproduced in the margin.)

Even though we do not know how to obtain the set of extreme coherent lower probabilities on some possibility space Ω analytically, or even if this is possible, there are some results we can deduce analytically.

(i) One of them is that this set includes the – evidently coherent – vacuous previsions relative to subsets $(1.38)_{43}$. To see this, assume that the vacuous probability $\underline{P}^A$ (i.e., relative to $A :\subseteq \Omega$) can be written as a convex combination of coherent lower probabilities:

$$\begin{aligned}&\exists \mathscr{Q} :\in (\mathscr{P}\mathscr{I}_\Omega)_{\mathrm{coh}};\\ &\exists \lambda : \Delta_{\mathscr{Q}};\\ &\qquad \textstyle\sum_{\underline{R}:\mathscr{Q}} \lambda_{\underline{R}} \cdot \underline{R} = \underline{P}^A,\end{aligned}$$

then, due to the $[0,1]$-boundedness of coherent probabilities, it must hold for all events B of Ω and all lower probabilities $\underline{R} : \mathscr{Q}$ that $\underline{R}B = \underline{P}^A B$. This means $\mathscr{Q} = \imath \underline{P}^A$.

(ii) Another result is that the degenerate probabilities are the only extreme coherent lower probabilities that are nonzero on some elementary event. Using normedness and superadditivity, this follows from the fact that any extreme coherent lower probability must be $\{0,1\}$-valued on singletons, which itself can be seen to hold using the following argument: Consider, ex absurdo, that there is a $\underline{P} : \mathrm{ext}(\mathscr{P}\mathscr{I}_\Omega)_{\mathrm{coh}}$ and an $\omega : \Omega$ such that $\underline{P}\omega \in]0,1[$. For every $P : \mathscr{M}\underline{P}$, $\frac{P - \underline{P}\omega \cdot P^\omega}{1-\underline{P}\omega}$ is an additive probability (we have just moved the probability mass $\frac{\underline{P}\omega}{1-\underline{P}\omega}$ away from ω to the other elementary events). So

$$\underline{R} := \min_{P:\mathrm{ext}(\mathscr{M}\underline{P})} \tfrac{1}{1-\underline{P}\omega} \cdot (P - \underline{P}\omega \cdot P^\omega),$$

is a coherent lower probability. Moving first the scaling factor $1 - \underline{P}\omega$ and then the term $\underline{P}\omega \cdot P^\omega$ to the left-hand side, we see that we can write $\underline{P} = \min_{P:\mathrm{ext}(\mathscr{M}\underline{P})} P = (1 - \underline{P}\omega) \cdot \underline{R} + \underline{P}\omega \cdot P^\omega$, contradicting the assumption that $\underline{P}$ is an extreme point. (Our proof here is based on an argument by Sebastian Maaß [2005].)

(iii) Because the labeling of the elements of Ω is irrelevant, the set of (extreme) coherent lower probabilities is invariant under permutation of the labeling. Adding requirements besides coherence may break this symmetry, of course.

(iv) Every coherent extreme lower probability on a possibility space Ω can be extended to an extreme coherent lower probability on a larger possibility space Ω'. This becomes clear by realizing that any coherent lower probability $\underline{P}$ on $\wp\Omega'$ such that $\underline{P}(\Omega' \setminus \Omega) = 0$ can only be written as a convex combination of extreme points for which the same holds. Each of these extreme points is therefore completely determined by their values on $\wp\Omega$.

Concerning k-monotonicity, there is one analytical result we must mention: k-monotonicity for any $k \geq |\Omega|$ is equivalent to complete monotonicity [Chateauneuf & Jaffray 1989, Corollary 1]. Our results do not contradict this, we always find the vacuous probabilities relative to events as extreme points in these cases (cf. $(2.8)_{72}$).

A last general remark concerns the combination of properties. When we look at the convex sets of lower probabilities satisfying different linear constraint-based properties, their intersection is the convex set of lower probabilities satisfying all those properties. Every extreme point of any of the original sets that also satisfies the other properties – i.e., lies in the intersection – is thus an extreme point of this intersection. A useful application of this insight is the following: We have mentioned (after $(2.10)_{73}$) that, for $k:\mathbb{N}_{\geq 2}$, nonnegative normed k-monotone lower probabilities are coherent. The vacuous lower probabilities relative to events are k-monotone by definition (cf. the two paragraphs before $(2.10)_{73}$) and are, as seen above in item $(\mathrm{i})_{\cap}$, extreme coherent lower probabilities. Therefore, the vacuous lower probabilities relative to events are extreme nonnegative normed k-monotone lower probabilities.

After these general remarks, we take a look at some more practical results in the next subsections. There, we only look at possibility spaces of three, four and six elementary events. The smaller cardinalities become trivial, as the restrictions imposed at the beginning of this subsection give the same results as coherence and even complete monotonicity:

$|\Omega| = 0$: then $\Omega = \emptyset$ and $\wp\Omega = \{\emptyset\}$; $(\mathscr{P}\mathscr{I}_{\emptyset})_{\mathrm{nrm}}$ is empty, because normedness $(2.4)_{71}$ cannot be satisfied. (Admittedly, this case is absurd.)

$|\Omega| = 1$: then $\wp\Omega = \{\emptyset, \Omega\}$; $(\mathscr{P}\mathscr{I}_{\Omega})_{\mathrm{nrm}}$'s single extreme point is the degenerate vacuous prevision $\underline{P}^{\Omega}$. (This case is useless if not absurd.)

For $|\Omega| = 5$, we found – after months of vertex enumeration by a computer – 1 743 093 of the extreme coherent lower probabilities; a hardware failure stopped us from finding the rest.

$|\Omega| = 2$: let $\Omega := \{a, b\}$; the three extreme points of $(\mathscr{P}\mathscr{I}_{\{a,b\}})_{\mathrm{nrm}\wedge\mathrm{nng}\wedge\mathrm{asl}}$ are the vacuous probability $\underline{P}^{\Omega}$ and the degenerate probabilities P^a and P^b. So in this first nonabsurd and useful case – often used in examples – all coherent probabilities are linear-vacuous, which makes it ill-suited to illustrate the peculiarities of coherence.

Due to the combinatorial explosion in the number of constraints for the interesting properties and the computational complexity of the vertex enumeration algorithms (cf. §2.1.3$_{68}$), results for possibility spaces larger than four are hard to compute.

2.3.2 *Three elementary events*

In this subsection, we let $\Omega := \{a, b, c\}$ and use the cardinality-then-lexicographical ordering of events; as an illustration, let us see what this gives for the power set and some special lower probabilities:

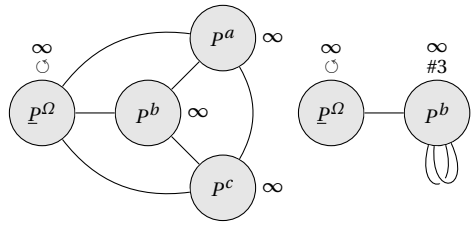
$$\begin{aligned}
\wp\Omega &= \{\emptyset,\ \imath a, \imath b, \imath c,\ \{a,b\},\{a,c\},\{b,c\},\ \Omega\},\\
\underline{P}^{\Omega} &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1),\\
\underline{P}^{\{a,b\}} &= (0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1),\\
P^{a} &= (0,\ 1,\ 0,\ 0,\ 1,\ 1,\ 0,\ 1).
\end{aligned}$$

Furthermore, it is economical, notation-wise, to write the lower probabilities more compactly without parentheses and commas as, for example,

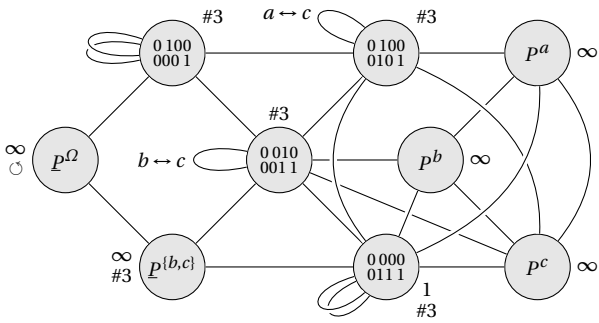
$P^a = 0\,100\,110\,1$, grouped per cardinality of the event. Stacked versions are also used; e.g., $P^a = \begin{smallmatrix} 0\,100 \\ 110\,1 \end{smallmatrix}$.

As an introductory example, we are going to look at the extreme linear-vacuous lower probabilities, or, more explicitly, at $\{\underline{P}^\Omega, P^a, P^b, P^c\}$. Apart from giving this set of extreme points, we can also give the corresponding adjacency graph, in which the nodes correspond to the extreme points and in which the edges connect neighboring extreme points of the polytope. The adjacency graph of the set defined by convex combinations of P^a, P^b, P^c, and $\underline{P}^\Omega$ is given by the left-hand side figure. The annotations (best ignored on a first reading) mean the following: ∞ indicates that the extreme lower probability is completely monotone (cf. §2.2.3$_{72}$); the circular arrow ↻ indicates that the extreme lower probability is permutation invariant, i.e., satisfies (2.27)$_{84}$. For the current case, the adjacency graph can be drawn in its entirety; this would result in too space-consuming spaghetti-like drawings for others. In some cases, this can be resolved by looking at the permutational symmetries in the graph: we can restrict ourselves to a partial graph that generates the whole graph – i.e., all of its nodes and edges – by looking at all permuted versions. This is shown in the right-hand side figure. In such a partial graph, some nodes represent a permutation class, a permutation-invariant subset of extreme lower probabilities; here, P^b represents $\{P^a, P^b, P^c\}$. To indicate edges within this subset, loops are used. To keep track of the total number of extreme lower probabilities, a #-prefixed number is added as annotation to indicate the number of extreme lower probabilities represented by the node.

The list of neighbors can be obtained with a vertex enumeration program [Fukuda & Prudon 1996].

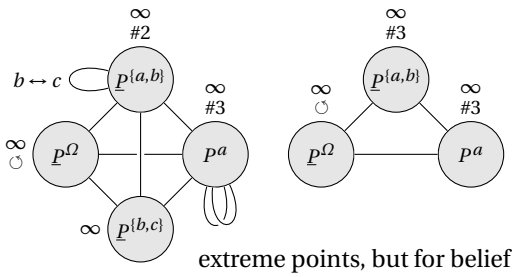


The first nontrivial case we are going to look at is lower probabilities that avoid sure loss, the set $(\mathscr{P}\mathscr{I}_\Omega)_{\text{asl}\wedge\text{nng}\wedge\text{nrm}}$, to be precise. This set has 19 extreme points; the corresponding adjacency graph is given in the left-hand side drawing, its permutationally restricted but legible counterpart is given on the right-hand side.



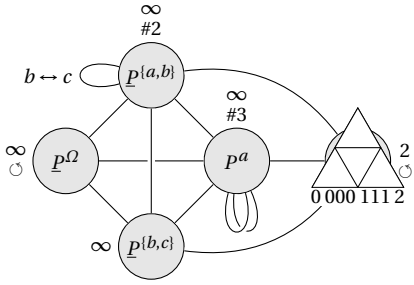
A first remark concerns the full graph: we have included it here just this once as a justification for not giving it again for any of the cases still to come. Some additional remarks about the restricted graph are also in order: When an extreme lower probability is nonmonotone, we omit the (undefined) monotonicity level. Some extreme lower probabilities, although related by permutation, are nevertheless given as separate nodes (e.g., $\begin{smallmatrix}0&0&1&0\\0&0&1&1\end{smallmatrix}$ and $\begin{smallmatrix}0&1&0&0\\0&1&0&1\end{smallmatrix}$); this is because they take up a different role in the restricted graph. This also forces us to indicate – by notationally abusing the bijection arrow $\leftrightarrow$ – to which permutation the loops on these nodes correspond.

After the relatively complicated case above, we turn to two cases with simpler adjacency graphs: We compare the graphs for the completely monotone belief functions $(\mathscr{P}\mathscr{I}_\Omega)_{\text{bel}}$ (cf. $(2.8)_{72}$; it is also equal to $(\mathscr{P}\mathscr{I}_\Omega)_{3\text{-mon}\wedge\text{nrm}}$) and the necessity measures $(\mathscr{P}\mathscr{I}_\Omega)_{\text{nec}\wedge\text{nrm}}$ (cf. §2.2.7; a subset of the belief functions). It is known that both sets have the vacuous previsions relative to events as extreme points, but for belief functions, all extreme points are neighbors (see left-hand side figure), for necessity measures, the 'extreme points' are grouped per lobe (see right-hand side figure). For $(\mathscr{P}\mathscr{I}_\Omega)_{\text{bel}}$, notice the need to create a separate node $\underline{P}^{\{b,c\}}$; it would otherwise not have been possible to express that it also neighbors P^a.



To finish this subsection with the most important case, we now look at coherence. As was also shown by Maaß [2003*a*, 2005], the set of eight extreme coherent lower probabilities on a possibility space of three elements is

$$\operatorname{ext}(\mathscr{P}\mathscr{I}_\Omega)_{\text{coh}} = \left\{P^a, P^b, P^c, \underline{P}^{\{a,b\}}, \underline{P}^{\{a,c\}}, \underline{P}^{\{b,c\}}, \underline{P}^{\Omega}, 0\,000\,\tfrac{1}{2}\tfrac{1}{2}\tfrac{1}{2}\,1\right\}.$$



This set consists of all the vacuous lower probabilities relative to events and also one new extreme point, which is 2-monotone. This implies – as k-monotonicity is preserved under convex combinations – that all lower probabilities on a possibility space of three elements are 2-monotone, so this is also the graph corresponding to $(\mathscr{P}\mathscr{I}_\Omega)_{2\text{-mon}\wedge\text{nrm}}$. In the adjacency graph we have added a graphical representation of the credal set of $0\,000\,\frac{1}{2}\frac{1}{2}\frac{1}{2}\,1$ (the credal sets of the vacuous lower probabilities are the 0..2-dimensional faces of the unit simplex). Due to their construction, all the extreme lower probabilities have rational components, so after rescaling, they can be written more tersely using only integers, as is illustrated in the figure.

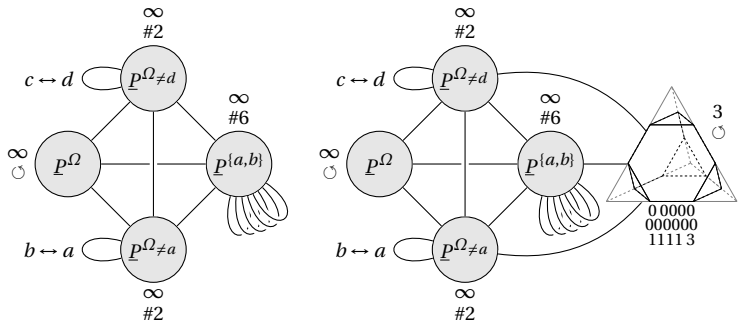
2.3.3 *Four elementary events*

In this subsection, we let $\Omega := \{a,b,c,d\}$; the power set with its ordering and some illustrative vacuous lower probabilities become

$$\begin{aligned}
\wp\Omega &= \{\emptyset,\ \imath a,\imath b,\imath c,\imath d,\ \{a,b\},\{a,c\},\{a,d\}\{b,c\},\{b,d\},\{c,d\},\ \Omega_{\neq d},\Omega_{\neq c},\Omega_{\neq b},\Omega_{\neq a},\ \Omega\},\\
\underline{P}^{\Omega} &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1),\\
\underline{P}^{\Omega_{\neq d}} &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 1),\\
\underline{P}^{\{a,b\}} &= (0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 1),\\
P^{a} &= (0,\ 1,\ 0,\ 0,\ 0,\ 1,\ 1,\ 1,\ 0,\ 0,\ 0,\ 1,\ 1,\ 1,\ 0,\ 1).
\end{aligned}$$

With the increase in size of the possibility space, and the corresponding doubling of the size of its power set, the adjacency graphs risk becoming unreadable (due to the increase in the number of nodes and connections), even when exploiting permutational symmetries. There is, however, a possibility for a further simplification of the graphs: We have observed that whenever coherence is implied by the properties studied, every extreme lower probability that is additive is connected to all other nodes (and I conjecture that this is not limited to the cases we studied). So we can omit these additive probabilities – the degenerate ones, in fact – from the two graphs below.

An illustration of this approach is given with the partial graph for the set of belief functions $(\mathscr{P}\mathscr{I}_\Omega)_{\text{bel}}$ on the left-hand side, and – on the right-hand side – with the partial graph for the set of 3-monotone lower probabilities $(\mathscr{P}\mathscr{I}_\Omega)_{3\text{-mon}\wedge\text{nrm}}$.Notice the parallels with what we have seen in the previous subsection.



In §2.1.3[68], we have already mentioned that the decomposition of a lower probability as a convex combination of extreme lower probabilities is not necessarily unique. The last case allows us to illustrate this fact nicely. The permutation invariant completely monotone lower probability 0 0000 000000 $\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}$ 1 can be written as two different convex combinations of extreme 3-monotone lower probabilities:

$$\begin{aligned}
0\,0000\,000000\ \tfrac{1}{4}\tfrac{1}{4}\tfrac{1}{4}\tfrac{1}{4}\ 1 &= \tfrac{3}{4}\cdot(0\,0000\,000000\ \tfrac{1}{3}\tfrac{1}{3}\tfrac{1}{3}\tfrac{1}{3}\ 1) + \tfrac{1}{4}\cdot\underline{P}^{\Omega}\\
&= \textstyle\sum_{\omega:\Omega}\tfrac{1}{4}\cdot\underline{P}^{\Omega_{\neq\omega}}.
\end{aligned}$$

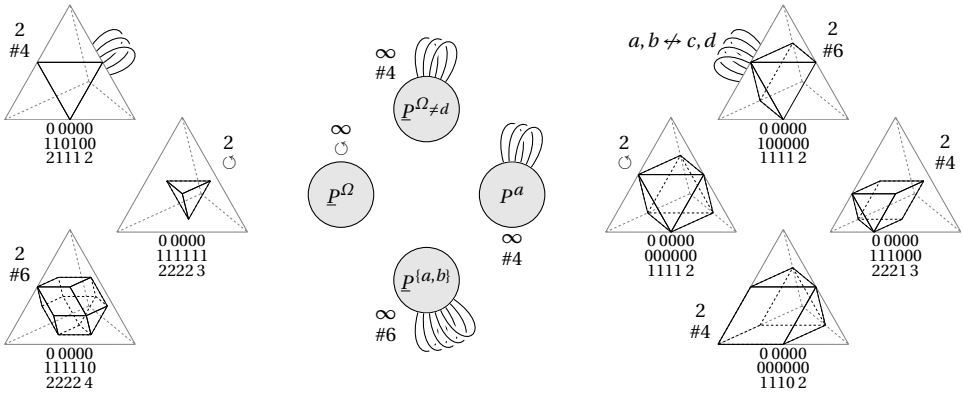
More generally, we can say the following: A polytope for which the adja-

cency graph is complete – i.e., for which all nodes are connected to each other – corresponds to a simplex, so any of its points has a unique decomposition in terms of extreme points (the left-hand side graph above is an example). Whenever some nodes are not connected (such as $\underline{P}^{\Omega}$ and 0 0000 000000 $\frac{1}{3}\frac{1}{3}\frac{1}{3}\frac{1}{3}$ 1 in the last case), Radon's theorem [Peterson 1972] tells us that the decomposition is not unique.

Looking at the adjacency graph for the 3-monotone lower probabilities on a possibility space of cardinality 4 (above) and at the one for 2-monotone lower probabilities on a possibility space of cardinality 3 (at the end of the previous subsection), we see a very similar structure; in comparison to the belief functions there is one extra extreme lower probability (characterized by the value $1/|\Omega|-1$ on events of cardinality $|\Omega|-1$) connected to all but the vacuous lower probability. The same structure is seen for a cardinality of 5, which makes me bold enough to conjecture that this might hold for all finite cardinalities.

In the end, there are always cases for which the adjacency graphs become too complex to draw even when leaving out the degenerate ones. It can nevertheless be interesting to give an ordered list of the extreme lower probabilities in those cases. Below, we do just that for the set $(\mathscr{P}\mathscr{I}_{\Omega})_{k\text{-mon}\wedge\text{nrm}}$ of 41 extreme 2-monotone lower probabilities. We still use the 'one example per permutation class'-simplification and have ordered the nodes such that those that are connected to the nodes for the vacuous lower probability are on the left of the vacuous lower probabilities and those that are not, on the right.

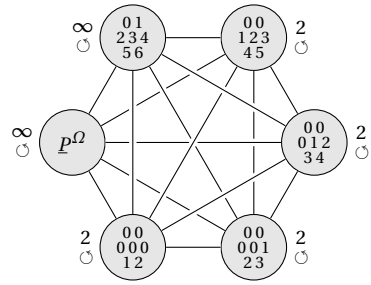
The extreme 2-monotone lower probabilities were already known about forty years ago [Shapley 1971].



The list of the most interesting set of extreme lower probabilities, the extreme coherent ones, is rather large, so we have relegated it to §A.1.1$_{194}$ in the Herbarium$_{194}$. There, you can also verify that amongst the extreme 2-monotone lower probabilities given above, only the ones in the lower-left permutation class (with 0 0000 $\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}0$ $\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$ 1 as a representative) are not extreme coherent lower probabilities.

2.3.4 *Staring two-monotonely at cubical dice*

For possibility spaces of five elements and larger, the computation of the extreme lower probabilities using our approach becomes too computationally intensive in many cases. The one we present here to draw the curtains over this chapter is amongst the ones that are still practical.



Let Ω be a six-element space, we then look at the set $(\mathscr{P}\mathscr{I}_{\Omega})_{2\text{-mon}\wedge\text{nrm}\wedge\text{pin}}$ of 2-monotone permutation invariant lower probabilities. This set contains models for the uncertainty faced when betting with common six-faced cubical dice for which no face-specific knowledge is available. As permutation invariant lower probabilities must be constant on events with constant cardinality (cf. §2.2.6$_{83}$), we have given only one component value per cardinality size when writing down the extreme lower probabilities in the nodes of the adjacency graph. (Note the regularity of the component values. The same regularity is also observed in §A.1.2$_{196}$ for smaller possibility spaces, so this makes for a nice conjecture.)

A fair die is modeled by $0\ \frac{1}{6}\ \frac{1}{3}\ \frac{1}{2}\ \frac{2}{3}\ \frac{5}{6}\ 1$, one we know nothing about by $\underline{P}^{\Omega}$, and $0\,0\,0\,0\,0\ \frac{1}{2}\ 1$ should be used when we only know that none of the die's faces is more likely to come up than the others combined.

Note that the monotonicity levels of the non-vacuous extreme lower probabilities had to be calculated, as it is a priori possible for some (but not all) of them to be more than 2-monotone (e.g., 3-monotone). This can be done quite easily by applying the Möbius transform (cf. (2.9)$_{73}$) and a result by Chateauneuf & Jaffray [1989, Proposition 4].

To be honest, I would have liked to call this subsection "Staring coherently at cubical dice", but trying to calculate the extreme permutation invariant coherent lower probabilities on a possibility space of cardinality 6 turned out to be a computational bridge too far for the current implementation and personal computers. For lower cardinality possibility spaces, there was no problem. For those who like to see more results, we have therefore included the adjacency graphs of all permutation invariant k-monotone or coherent lower probabilities for these cases in the Herbarium$_{194}$, in §A.1.2$_{196}$.

INFERENCE MODELS

> Let men be once fully perswaded of these two principles, *that there is nothing in any object, consider'd in itself, which can afford us a reason for drawing a conclusion beyond it;* and, *that even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience;* I say, let men be once fully convinc'd of these two principles, and this will throw them so loose from all common systems, that they will make no difficulty in receiving any, which may appear the most extraordinary.
>
> Hume [1739, §1.3.12, ¶20]

We wish to learn from samples: make predictions about future samples or draw conclusions about the process generating the samples. This is respectively called predictive and parametric inference. The uncertainty model making these inferences based on the given samples is called an inference model.

The theory introduced in 'Modeling uncertainty'$_{28}$ allows us – using natural or regular extension – to make explicit what is implicit in the coherent lower previsions that describe what we know about some situation or problem. This is a form of deductive inference.

Howson [2000] argues that an epistemic probability theory can be used as a logic of induction. He stresses, however, that it must start with some (by that theory) unjustified premisses.

It would be nice to be able to draw definite inferences using only this and an observed sample. We can: the theory predicts that anything is possible and it tells us to conclude that we know nothing [Walley 1991, §7.3.7$_{366}$]. It would be even nicer to be able to obtain non-vacuous inferences. For this, at least a little information about the process generating the samples (we wish to learn about) must be added in some way. This is possible by making a leap of faith and adding a limited number of predictions or conclusions ourselves (e.g., by elicitation). Typically, this takes the form of a so-called *prior* uncertainty model that is itself non-vacuous.

This leap of faith is an essential aspect of what learning – inductive inference – is about. The method of trial & error, also called guess & check, is a prime example of this. In this chapter and the next, we are going to present inference models that are meant to allow people to make educated but conservative guesses. We do not focus on the checking or validation aspect of learning.

The inference models we are going to look at more or less transform inductive inference to deductive inference by providing some prior uncertainty model, from which an updated or posterior uncertainty model is obtained by conditioning on the observation. This is assumed to always be a (partial) sample and nothing more, as other types of information cannot be identified with a conditioning event. The updated model is then taken to be the model learned from the sample. Note however, that even though the approach we are taking is comfortingly consistent, there is no reason to preclude other approaches, in which learning is not (directly) equated to conditioning. Walley [1991, §6.11$_{334}$] provides a nice basis for reflection on this issue, which is especially important whenever the observations cannot by good approximation be identified with some conditioning event.

In this chapter, the ideas just mentioned are made concrete by looking at inference models for categorical data, a term that refers to the fact that the considered sample spaces are finite and thus discrete. On the one hand, we look at a framework for immediate predictive inference in §3.2$_{118}$ that leads to the predictive inference model called the imprecise Dirichlet-multinomial model – IDMM for short – [Walley & Bernard 1999]. On the other hand, in §3.3$_{138}$, we look at the imprecise Dirichlet model – IDM for short – [Walley 1996], a parametric inference model. They are followed by some applications in §3.4$_{142}$. But first we strengthen our foundations by expounding on the concepts of exchangeability, sufficient statistics, and likelihood functions; this will prove essential for the ensuing sections.

Structural judgements [Walley 1991, Chapter 9$_{442}$] such as exchangeability play an important role in the design of inference models.

3.1 Exchangeability, sufficient statistics & likelihood functions

In this section, we do the basic set-up for the inference models we consider. We first look at the exchangeability assumption and its consequences; it constitutes an essential ingredient in our approaches. This discussion leads to the introduction of some other concepts, such as sufficient statistics and likelihood functions.

The consequences of a slightly different definition of exchangeability are investigated by De Cooman et al. [2007*b*, 2009*c*].

3.1.1 *Samples, random variables & exchangeability*

Consider a subject who is making a finite sequence of $N : \mathbb{N}_{>0}$ observations of a certain phenomenon. This phenomenon can for example be the amount and type of precipitation on Christmas eve, the number of deadly car accidents every weekend, or the number and type of the defects in a freshly grown silicon crystal.

In this chapter, we only consider phenomena that provide observations that can be put in a finite number of categories; the archetypical example for this is drawing colored marbles from an urn. We call each observation a sample and the set of possible observations (categories)

the sample space, generically denoted by $\mathscr{X}$. This results in a possibility space $\mathscr{X}^N$, i.e., the space of possible sequences.

As we assume that the same sample space $\mathscr{X}$ can be used for each observation in the sequence, we can use a sequence of random variables $X := (X_k \mid k : 1..N)$ to be able to clearly identify each of the observations and avoid confusion about what is observed when. Each of these random variables is just the identity function on $\mathscr{X}$; so, when observation number $\ell : 1..N$ turns out to be the category z, then $X_\ell z = z$; or $X_\ell = z$ for short.

The beliefs the subject has about these random variables – i.e., about how likely it is to observe some sequence of values – can be modeled with a coherent set of desirable gambles $\mathscr{R} :\subset \mathscr{L}_{\mathscr{X}^N}$. We call the corresponding unconditional coherent lower prevision $\underline{P}$ on $\mathscr{L}_{\mathscr{X}^N}$ the (joint) sequence distribution of X. They are related by $\mathscr{G}_{\underline{P}} = \mathscr{G}_{\mathscr{R}}$ (cf. $(1.11)_{36}$, $(1.12)_{36}$, and $(1.19)_{37}$).

When the subject has reasons to believe the process that generates the samples is the same, practically speaking, for every observation, then the order of observation is irrelevant and the subject should make no practical distinction between a sequence of random variables with or without permuted indices. To formalize what this means for $\mathscr{R}$ and $\underline{P}$, we lift permutations from the index set to vectors of samples and to gambles:

Illustration with $\mathscr{X} := \{a, b\}$, $N := 3$, $x := a, b, a$, and $\pi := 3, 1, 2$: $\pi x = a, a, b$.

$$\Pi_{1..N} := \left\{ \pi : \begin{array}{ccc} 1..N & \leftrightarrow & 1..N \\ \cup\, \mathscr{X}^N & & \cup\, \mathscr{X}^N \\ \cup\, \mathscr{L}_{\mathscr{X}^N} & & \cup\, \mathscr{L}_{\mathscr{X}^N} \end{array} \middle| \begin{array}{l} \forall x : \mathscr{X}^N ; \pi x = x \circ \pi \\ \wedge \forall f : \mathscr{L}_{\mathscr{X}^N} ; \pi f = f \circ \pi \end{array} \right\}. \tag{3.1}$$

Let π be any permutation in $\Pi_{1..N}$. *The belief that the phenomenon under study is invariant under the permutation of observations* – what we will from now on call exchangeability – *implies that the subject is indifferent between any gamble f on $\mathscr{X}^N$ and its permutations πf: he is marginally willing to exchange one for the other.* This means that

Walley [1991, §4.1.1$_{169}$] lists a number of direct judgements (among them indifference).

$$\underline{P}(\pi f - f) = \overline{P}(\pi f - f) = 0 = \overline{P}(f - \pi f) = \underline{P}(f - \pi f). \tag{3.2}$$

Define

$$\mathscr{H}\mathscr{X}^N := \operatorname{span}\{\pi f - f \mid f : \mathscr{L}_{\mathscr{X}^N} ; \pi : \Pi_{1..N}\}. \tag{3.3}$$

It is interesting to realize that the linear subspace $\mathscr{H}\mathscr{X}^N$ of $\mathscr{L}_{\mathscr{X}^N}$ only contains marginally desirable gambles, i.e., $\mathscr{H}\mathscr{X}^N \subseteq \mathscr{G}_{\underline{P}} = \mathscr{G}_{\mathscr{R}}$. To show this, consider gambles f and g on $\mathscr{X}^N$, two index permutations π and $\bar{\pi}$, a positive real number α, and a negative real number β, then

$$\underline{P}\big(\alpha \cdot (\pi f - f)\big) = \alpha \cdot \underline{P}(\pi f - f) = 0 = \beta \cdot \overline{P}(\bar{\pi} g - g) = \underline{P}\big(\beta \cdot (\bar{\pi} g - g)\big),$$
$$\underline{P}\big((\pi f - f) + (\bar{\pi} g - g)\big) \geq \underline{P}(\pi f - f) + \underline{P}(\bar{\pi} g - g) = 0,$$

$$\underline{P}\big((\pi f - f) + (\bar{\pi} g - g)\big) \leq \underline{P}(\pi f - f) + \bar{P}(\bar{\pi} g - g) = 0,$$

where we have used (3.2), nonnegative homogeneity $(1.30)_{42}$, conjugacy $(1.14)_{36}$, superadditivity $(1.31)_{42}$, and mixed subadditivity $(1.36)_{42}$. By inverting the inequalities and interchanging $\underline{P}$ and $\bar{P}$ we also obtain a set of valid expressions. This means that $\underline{P}$ is both zero *and* self-conjugate, and therefore linear, on the whole of $\mathscr{H}\mathscr{X}^N$.

Knowing that a coherent lower prevision $\underline{P}$ on $\mathscr{X}^N$ is completely defined by its set of marginal gambles, the realization that $\mathscr{H}\mathscr{X}^N \subseteq \mathscr{G}_{\underline{P}}$ under an exchangeability assumption is used as a definition of exchangeability for lower previsions. The predicate $\mathrm{xch}:(\mathscr{P}\mathscr{L}_{\mathscr{X}^N})_{\mathrm{coh}} \to \mathbb{B}$ formalizes this:

$$\begin{aligned}\mathrm{xch}\,\underline{P} \Leftrightarrow \mathscr{H}\mathscr{X}^N \subseteq \mathscr{G}_{\underline{P}} \Leftrightarrow \forall f : \mathscr{L}_{\mathscr{X}^N};\\ \forall \pi : \Pi_{1..N};\\ \underline{P}(\pi f - f) = 0 = \bar{P}(\pi f - f).\end{aligned} \tag{3.4}$$

On the other hand, $\mathscr{H}\mathscr{X}^N \subseteq \mathscr{G}_{\mathscr{R}}$ cannot serve as a definition of exchangeability for coherent sets of desirable gambles $\mathscr{R} :\subset \mathscr{L}_{\mathscr{X}^N}$, considering that $\mathscr{G}_{\mathscr{R}}$ cannot encode the borderline behavior of $\mathscr{R}$; it can serve as a basis, however. As any gamble in $\mathscr{H}\mathscr{X}^N$ is marginally desirable, the sum of such a gamble and a desirable gamble is also desirable, as long as it avoids partial loss. We formalize this as

$$\mathscr{H}\mathscr{X}^N + \mathscr{R}_{\neq(\mathscr{X}^N;0)} \subseteq \mathscr{R}, \tag{3.5}$$

The least committal exchangeable coherent set of desirable gambles is $\mathscr{H}\mathscr{X}^N + (\mathscr{L}_{\mathscr{X}^N})_{\geq 0 \wedge \neq(\mathscr{X}^N;0)}$.

and take this constraint on $\mathscr{R}$ as a definition of exchangeability for coherent sets of desirable gambles. We use $\mathscr{R}_{\neq(\mathscr{X}^N;0)}$ instead of $\mathscr{R}$ on the left-hand side to avoid letting the desirability of gambles in $\mathscr{H}\mathscr{X}^N$ depend on the desirability of the zero gamble $(\mathscr{X}^N;0)$, which seems just a question of terminology. We do not have to worry about the left-hand side not avoiding partial loss, because $\mathscr{H}\mathscr{X}^N$ contains no gambles leading to a partial loss; i.e., whenever $(\pi f - f)x < 0$ for some sequence x, then

$$0 = \textstyle\sum(\pi f - f) < \sum_{y:\mathscr{X}^N \wedge y \neq x}(\pi f - f)y,$$

so there is some y in $\mathscr{X}^N$ different from x such that $(\pi f - f)y > 0$.

The left-hand side of (3.5) is very much akin to the construction of a set of desirable gambles starting from a set of marginally desirable gambles $\mathscr{G}_{\underline{P}}$ (of some lower prevision $\underline{P}$) we encountered in §1.2.2$_{36}$. We mentioned $\mathscr{G}_{\underline{P}} + \mathbb{R}_{>0}$ and $\mathscr{G}_{\underline{P}} + (\mathscr{L}_{\mathscr{X}^N})_{>0}$ as good options for the corresponding set of desirable gambles (cf. $(1.21)_{37}$ and the surrounding paragraph). Here, the first difference is that we only take a subset of the marginally desirable gambles, so we only generate a subset of the desirable gambles. The second difference is that we know all desirable gambles and

therefore do not need to restrict ourselves to $\mathbb{R}_{>0}$ or $(\mathscr{L}_{\mathscr{X}^N})_{>0}$. We will see further on (cf. $(3.48)_{110}$ and below) that thanks to this, $(3.5)_\cap$ also implies that after an observation, the updated sets of desirable gambles of exchangeable sets of desirable gambes are exchangeable as well.

We have given two definitions of exchangeability: one for previsions and one for sets of desirable gambles. If we wish to use both types of uncertainty model interchangeably, we need to show that they are consistent in the sense that the prevision corresponding to an exchangeable set of desirable gambles is exchangeable and that any exchangeable prevision can be derived from some exchangeable set of desirable gambles.

(i) We first start from an exchangeable coherent set of desirable gambles $\mathscr{R}$. Considering that
   (a) $(\mathscr{X}^N;0) \in \mathrm{cl}\,\mathscr{R}$, so $\mathscr{H}\mathscr{X}^N = \mathscr{H}\mathscr{X}^N + \iota(\mathscr{X}^N;0) \subset \mathscr{R} \cup \mathscr{G}_{\mathscr{R}}$, and
   (b) $\mathscr{H}\mathscr{X}^N$ is a linear subspace of $\mathscr{X}^N$ that contains the zero gamble $(\mathscr{X}^N;0)$, so it lies either entirely on the border $\mathscr{G}_{\mathscr{R}}$ of the cone that is $\mathscr{R}$ or part of it lies strictly outside of this border,

   we must conclude that $\mathscr{H}\mathscr{X}^N \subseteq \mathscr{G}_{\mathscr{R}}$. Thus the coherent lower prevision $\underline{P}$ corresponding to $\mathscr{R}$ satisfies $(3.4)_\cap$, because $\mathscr{G}_{\underline{P}} = \mathscr{G}_{\mathscr{R}}$.

(ii) Starting from an exchangeable coherent lower prevision $\underline{P}$, the corresponding least committal coherent set of desirable gambles $\mathscr{R}_{\underline{P}}$ (cf. $(1.22)_{38}$) might not satisfy $(3.5)_\cap$. However, the enlarged set $(\mathscr{R}_{\underline{P}})_{=(\mathscr{X}^N;0)} \cup \big(\mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)}\big)$ does:

$$\begin{aligned}
&\mathscr{H}\mathscr{X}^N + \Big((\mathscr{R}_{\underline{P}})_{=(\mathscr{X}^N;0)} \cup \big(\mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)}\big)\Big)_{\neq(\mathscr{X}^N;0)} \\
&\subseteq \mathscr{H}\mathscr{X}^N + \big(\mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)}\big);
\end{aligned}$$

because $\mathscr{H}\mathscr{X}^N$ is a linear space, this becomes

$$\begin{aligned}
&= \mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)} \\
&\subseteq (\mathscr{R}_{\underline{P}})_{=(\mathscr{X}^N;0)} \cup \big(\mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)}\big).
\end{aligned}$$

This enlarged set is coherent, because the set-sum of the linear, partial loss-avoiding, $(\mathscr{X}^N;0)$-containing set $\mathscr{H}\mathscr{X}^N$ and $(\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)}$ is a cone satisfying (1.6)–$(1.9)_{35}$. Moreover, $\mathscr{H}\mathscr{X}^N + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)} \subseteq \mathscr{G}_{\underline{P}} + (\mathscr{R}_{\underline{P}})_{\neq(\mathscr{X}^N;0)} \subseteq \mathscr{G}_{\underline{P}} \cup \mathscr{R}_{\underline{P}}$, so the enlarged set's marginal gambles are the same and therefore so is the corresponding lower prevision, $\underline{P}$.

We mainly work in terms of previsions in this chapter, but when needed, we use the stronger definition $(3.5)_\cap$ of exchangeability for desirable gambles.

The self-conjugacy of exchangeable previsions on gambles of the form $\pi f - f$ has some interesting immediate consequences:

(i) Vacuous previsions cannot be exchangeable: they are only self-conjugate on constant gambles.

(ii) All lower previsions that dominate an exchangeable lower prevision are also exchangeable; so its credal set (cf. §1.2.9$_{50}$) must consist of exchangeable linear previsions only:

$$\operatorname{xch}\underline{P} \Leftrightarrow \mathscr{M}\underline{P} = (\mathscr{M}\underline{P})_{\operatorname{xch}}. \tag{3.6}$$

(iii) We can infer from the identity in (3.4)$_{97}$, using mixed sub and superadditivity (1.36)$_{42}$, that

$$\underline{P}(\pi f) - \underline{P} f \geq 0 \geq \underline{P}(\pi f) - \underline{P} f,$$

so we then see that exchangeability implies permutation invariance or permutability (cf. §2.2.6$_{83}$):

$$\begin{aligned} \operatorname{xch}\underline{P} \Rightarrow \forall f : \mathscr{L}_{\mathscr{X}^N}; \\ \forall \pi : \Pi_{1..N}; \\ \underline{P}(\pi f) = \underline{P} f. \end{aligned} \tag{3.7}$$

The converse does not hold, as vacuous previsions are permutable.

When the sequence distribution $\underline{P}$ of the sequence of random variables X is exchangeable, we also call that sequence exchangeable. Consider this to be the case, then any permutation $(X_{\pi k} \mid k : 1..N)$ is exchangeable as well and has the same sequence distribution $\underline{P}$. Moreover, the vector formed by any selection of $n : 1..N$ of the random variables in X is also exchangeable. Its sequence distribution is the exchangeable $\mathscr{X}^n$-marginal $\underline{P}^n$ on $\mathscr{L}_{\mathscr{X}^n}$ of $\underline{P}$ (cf. §1.3.1$_{52}$).

In the previous chapter, properties imposed on lower probabilities restricted the form of these lower probabilities, i.e., they could be written as a convex combination of a specific set of extreme lower probabilities. Here, we also impose a property – exchangeability –, now on a lower prevision. It turns out that this also restricts the form of these lower previsions, but in a different way. This is the subject of the next subsection.

3.1.2 *Representation in terms of count vectors*

We have just seen that the order of an exchangeable sequence of random variables and thus of a finite sequence of samples x in $\mathscr{X}^* := \bigcup_{n:\mathbb{N}} \mathscr{X}^n$ is irrelevant. What remains of this length-νx sample sequence after discarding the order can be summarized by a vector $m := \left|\{k : 1..\nu x \mid x_k = z\}\right|$ in $\mathbb{N}^{\mathscr{X}}$ containing the number of times m_z each category z of $\mathscr{X}$ occurs in x. We call m the count vector of the sequence x. Of course it is also the count vector of any permutation πx of x.

To the space $\mathscr{X}^N$ of all sample sequences of length N, there corresponds a space of count vectors

$$N^{\mathscr{X}} := \left\{m : \mathbb{N}^{\mathscr{X}} \,\middle|\, \sum m = N\right\}. \tag{3.8}$$

The surjective counting map $C_{\mathscr{X}} : \mathscr{X}^* \to \mathbb{N}^{\mathscr{X}}$ that maps sequences to count vectors is defined for every sequence x and every category z of $\mathscr{X}$ by

$$(C_{\mathscr{X}} x)_z = \left|\{k : 1..\nu x \mid x_k = z\}\right|, \tag{3.9}$$

where νx is the length of the sample sequence x (which is N in this subsection).

Take $x := a, b, a$ and $\pi x := a, a, b$, then $m = m_a, m_b := C_{\mathscr{X}} x = C_{\mathscr{X}}(\pi x) = 2, 1.$

Naturally, some sample sequences have the same count vectors; to wit, all those which are related by permutation. Given some count vector m in $\mathbb{N}^{\mathscr{X}}$ of total size $\nu m := \sum m$ (which is N in this subsection), the corresponding set of sample vectors – called atom – is

$$[m] := \left\{y : \mathscr{X}^{\nu m} \,\middle|\, C_{\mathscr{X}} y = m\right\}. \tag{3.10}$$

$[(2,1)] := \{(a,a,b), (a,b,a), (b,a,a)\}.$

Similarly, the atom of a sequence x is

$$[x] := \{\pi x \mid \pi : \Pi_{1..\nu x}\}. \tag{3.11}$$

For convenience, we extend the definition of the counting map to atoms with $C_{\mathscr{X}}[m] := m$ and $C_{\mathscr{X}}[x] := C_{\mathscr{X}} x$. The atom's size follows from counting the number of permutations:

$C_{\mathscr{X}}[(2,1)] = 2, 1.$

$|[(2,1)]| = \binom{3}{2,1} = \frac{3!}{2!\cdot 1!} = 3.$

$$|[m]| = \binom{\nu m}{m} := \frac{\nu m!}{\prod_{z:\mathscr{X}} m_z!}. \tag{3.12}$$

The set of atoms

$$[N^{\mathscr{X}}] := \left\{[m] \,\middle|\, m : N^{\mathscr{X}}\right\} \tag{3.13}$$

$[N^{\mathscr{X}}] := \{[(3,0)], [(2,1)], [(1,2)], [(0,3)]\}.$

is the partition of $\mathscr{X}^N$ generated by all permutations.

With any gamble h on $N^{\mathscr{X}}$ we can associate the gamble

$$h \circ C_{\mathscr{X}} = \textstyle\sum_{m:N^{\mathscr{X}}} h(m) \cdot I^{[m]} \tag{3.14}$$

on $\mathscr{X}^N$. Note that the function $(\circ C_{\mathscr{X}})$ is linear, constant additive, and supremum-preserving; also note that $\pi(h \circ C_{\mathscr{X}}) = h \circ C_{\mathscr{X}}$ for any permutation π.

We have now gathered all the ingredients we need to define the coherent lower prevision $\underline{Q}$ on $\mathscr{L}_{N^{\mathscr{X}}}$ induced by some exchangeable lower prevision $\underline{P}$ on $\mathscr{L}_{\mathscr{X}^N}$ (cf. §1.3.1$_{52}$). For every gamble h on $N^{\mathscr{X}}$, this prevision is defined by

$$\underline{Q} h = \underline{P}(h \circ C_{\mathscr{X}}). \tag{3.15}$$

This so-called count distribution is a model for the uncertainty about the count random variable $C_{\mathscr{X}} X$; it can be seen as a version of $\underline{P}$ from which exchangeability is abstracted away by using count vectors.

So what does this essence of exchangeability, which we have abstracted away, look like?

To find that out, we return to the basis of our definition of exchangeable sets of desirable gambles: somewhat below (3.3)$_{96}$, we found that $\underline{P}$

is zero and linear on $\mathscr{H}\mathscr{X}^N = \text{span}\{\pi f - f | f : \mathscr{L}_{\mathscr{X}^N}; \pi : \Pi_{1..N}\}$. Now fix some f in $\mathscr{L}_{\mathscr{X}^N}$, then $\underline{P}$ is zero and linear on $\text{span}\{\pi f - f \mid \pi : \Pi_{1..N}\}$; by focusing on the mean over this set we get

$$\underline{P}\big(\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f - f\big) = 0 = \overline{P}\big(\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f - f\big)$$

applying mixed sub- and superadditivity $(1.36)_{42}$ gives

$$\underline{P}\big(\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f\big) - \underline{P} f \geq 0 \geq \underline{P}\big(\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f\big) - \underline{P} f.$$

So

$$\underline{P} f = \underline{P}\big(\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f\big), \tag{3.16}$$

which means that the lower prevision of a gamble is equal to the prevision of its mean over all permutations, a result that is very similar to the permutability $(3.7)_{99}$ of an exchangeable lower prevision.

The full implications of this last equality become clear after an analysis of the right-hand side gamble. It is $[N^{\mathscr{X}}]$-measurable, i.e., constant on the atoms of the partition of $\mathscr{X}^N$ generated by the set of all permutations: let x be some sequence of samples, the corresponding atom is $[x] = \{\bar{\pi} x \mid \bar{\pi} : \Pi_{1..N}\}$ (cf. (3.11)). So if $\bar{\pi}$ is some index permutation, then

$$\textstyle\sum_{\pi:\Pi_{1..N}}(\pi f)(\bar{\pi} x) = \sum_{\pi:\Pi_{1..N}}\big((\bar{\pi}\circ\pi) f\big)x = \sum_{\pi:\Pi_{1..N}}(\pi f)x,$$

as the set of all permutations is invariant under permutation. Knowing this, we can rewrite this gamble in a way that makes its $[N^{\mathscr{X}}]$-measurability explicit:

$$\begin{aligned}
\tfrac{1}{N!}\cdot\textstyle\sum_{\pi:\Pi_{1..N}}\pi f &= \textstyle\sum_{m:N^{\mathscr{X}}}\frac{1}{|[m]|}\cdot\big(\sum_{x:[m]}\frac{1}{N!}\cdot\sum_{\pi:\Pi_{1..N}}(\pi f)x\big)\cdot I^{[m]} \\
&= \textstyle\sum_{m:N^{\mathscr{X}}}\frac{1}{|[m]|}\cdot\big(\frac{1}{N!}\cdot\sum_{\pi:\Pi_{1..N}}\sum_{x:[m]} f(\pi x)\big)\cdot I^{[m]} \\
&= \textstyle\sum_{m:N^{\mathscr{X}}}\frac{1}{|[m]|}\cdot\big(\frac{1}{N!}\cdot\sum_{\pi:\Pi_{1..N}}\sum_{x:[m]} f x\big)\cdot I^{[m]} \\
&= \textstyle\sum_{m:N^{\mathscr{X}}}\big(\frac{1}{|[m]|}\cdot\sum_{x:[m]} f x\big)\cdot I^{[m]} \\
&= \textstyle\sum_{m:N^{\mathscr{X}}}\text{Mh}(f|m)\cdot I^{[m]} = \text{Mh}(f|C_{\mathscr{X}}{}_{\bullet}),
\end{aligned} \tag{3.17}$$

where $\text{Mh}(\bullet|m)$ is the exchangeable linear prevision associated with the uniform distribution on $[m]$ defined by (let g be a gamble on $[m]$)

$$\text{Mh}(g|m) \coloneqq \tfrac{1}{|[m]|}\cdot\textstyle\sum_{y:[m]} g y. \tag{3.18}$$

According to our notational convention, $\text{Mh}(f|m) = \text{Mh}(f_{[m]}|m)$ (cf. §1.3.2$_{54}$).

This distribution is known as the multivariate hypergeometric distribution [Johnson et al. 1997, §39.2$_{171}$], which – not surprisingly – gives the probability of drawing a sequence x *without replacement* from an urn with composition m (or $y : \mathscr{X}^n$, when including its marginals for sequences of length $n : 1..N$).

Combining (3.16), (3.17), and (3.15) leads to a nice representation theorem in the spirit of the one for linear previsions by de Finetti [1937], but obtained directly for the more general coherent lower previsions.

Cifarelli & Regazzini [1996] provide an overview of de Finetti's contribution to probability theory.

It shows that the exchangeability of a sequence of random variables X or its distribution, the lower prevision $\underline{P}$ is intimately linked to count vectors and draws from an urn without replacement:

$$\begin{aligned}\operatorname{xch}\underline{P} \Leftrightarrow \underline{P} &= \underline{Q}\big(\mathrm{Mh}(\bullet|\bullet)\big),\\ &\text{with } \underline{Q} := \underline{P}(\bullet \circ C_{\mathscr{X}}).\end{aligned} \tag{3.19}$$

The importance of this result is the following: Under an assumption of exchangeability, we can – and perhaps should, to avoid not taking into account exchangeability – think in terms of count vectors (using the count distribution $\underline{Q}$); any additional information given as supremum prices for gambles on $\mathscr{X}^N$ can be translated to information for gambles on $N^{\mathscr{X}}$ using (3.19).

We can interpret $\mathrm{Mh}(\bullet|C_{\mathscr{X}}\bullet)$ as a conditional linear prevision corresponding to the partition $[N^{\mathscr{X}}]$, defined on $\mathscr{L}_{\mathscr{X}^N}$. This way, the characterization of the sequence distribution $\underline{P}$ through (3.19) can also be seen as the marginal extension of the conditional $\mathrm{Mh}(\bullet|\bullet)$ expressing only exchangeability and a marginal $\underline{Q}$ expressing all other information contained in $\underline{P}$. An immediate consequence of this is that whenever $\underline{P}[m] = \underline{Q}m > 0$ for some count vector m, then $\mathrm{Mh}(\bullet|m)$ is the unique solution of the GBR $(1.82)_{61}$.

Each of the marginal sequence distributions $\underline{P}^n$ – also being exchangeable – has an induced coherent count distribution $\underline{Q}^n$ on $\mathscr{L}_{n^{\mathscr{X}}}$, which can be used in a representation similar to the one described in (3.19). Because the space $N^{\mathscr{X}}$ of count vectors does not have a cartesian product structure, these count distributions $\underline{Q}^n$ cannot be viewed as simple marginals of $\underline{Q}$; they can, however, be viewed as induced distributions (cf. §1.3.1$_{52}$).

The function $S_n^N : \mathscr{L}_{n^{\mathscr{X}}} \to \mathscr{L}_{N^{\mathscr{X}}}$ characterizing their relationship can be found by doing the marginalization work on the level of the sequence distributions $\underline{P}$ and $\underline{P}^n$: (let h be a gamble on $n^{\mathscr{X}}$ and $f := h \circ C_{\mathscr{X}}$)

$$\underline{Q}(S_n^N h) := \underline{Q}^n h = \underline{P}^n f = \underline{P}\tilde{f} = \underline{Q}\big(\mathrm{Mh}(\tilde{f}|\bullet)\big), \tag{3.20}$$

where we have respectively used the definition of the count distribution, cylindrical extension of f to $\tilde{f}$ to go from marginal to joint distribution (cf. §1.3.1$_{52}$), and finally the representation theorem (3.19). Let us make the definition of S_n^N more explicit; consider some m in $N^{\mathscr{X}}$, then

$$(S_n^N h)m = \mathrm{Mh}(\tilde{f}|m) = \tfrac{1}{|[m]|} \cdot \textstyle\sum_{x:[m]} \tilde{f}x;$$

any sequence x in $[m]$ can be written as a pair of sequences y, z in $\bigcup_{m':n^{\mathscr{X}} \wedge m' \leq m} [m'] \times [m - m']$, so we can expand the above into

$$= \tfrac{1}{|[m]|} \cdot \textstyle\sum_{m':n^{\mathscr{X}} \wedge m' \leq m} \sum_{z:[m-m']} \sum_{y:[m']} fy;$$

to continue, use $(3.14)_{100}$ and realize that $[m-m'] = \emptyset$ when $m' \not\leq m$:

$$= \tfrac{1}{|[m]|} \cdot \textstyle\sum_{m' : n^{\mathscr{X}}} |[m-m']| \cdot \sum_{y : [m']} hm'$$
$$= \textstyle\sum_{m' : n^{\mathscr{X}}} \frac{|[m-m']| \cdot |[m']|}{|[m]|} \cdot hm'. \tag{3.21}$$

This explicit definition will be of use in the next subsection, where we let $N = \nu m$ become arbitrarily large.

3.1.3 *Arbitrary length sample sequences & limiting frequencies*

When we consider an infinite sequence of samples, our sequence of random variables becomes $X := (X_k \mid k : \mathbb{N}_{>0})$. We extend the definition of exchangeability to such infinite sequences by requiring that any finite subsequence should be exchangeable, or, equivalently, that for any n in $\mathbb{N}_{>0}$ the sequence $(X_k \mid k : 1..n)$ should be exchangeable. The possibility space now becomes $\mathbb{N}_{>0} \to \mathscr{X}$.

De Cooman et al. [2009c] give a more extensive and detailed treatment of the ideas we wish to convey in this subsection.

We could model the available information about X with its sequence distribution, a coherent lower prevision $\underline{P}$ on $\mathscr{L}_{\mathbb{N}_{>0} \to \mathscr{X}}$. The definition of exchangeability for X implies a definition for the exchangeability of $\underline{P}$: for every n, its coherent $\mathscr{X}^n$-marginal $\underline{P}^n$ on $\mathscr{L}_{\mathscr{X}^n}$ has to be an exchangeable sequence distribution. Knowledge of the supremum buying price defined by some marginal $\underline{P}^n$ in some gamble f on $\mathscr{X}^n$, defines the value of $\underline{P}$ for the cylindrically extended gamble $\tilde{f}$ on $\mathbb{N}_{>0} \to \mathscr{X}$ with $\underline{P}\tilde{f} = \underline{P}^n f$.

We wish to characterize the exchangeability of sample sequences of arbitrary length. However, for this task the possibility space $\mathbb{N}_{>0} \to \mathscr{X}$ appears quite awkward to work with and therefore so does $\underline{P}$. This is why we eliminate $\underline{P}$ from consideration, but preserve its unawkward marginals $(\underline{P}^n \mid n : \mathbb{N}_{>0})$ and the information they carry. We can get more insight by looking at frequency distributions, which are just count distributions under a thin disguise. So first consider, for every n, the coherent count distribution $\underline{Q}^n$ on $\mathscr{L}_{n^{\mathscr{X}}}$ corresponding to the marginal sequence distribution $\underline{P}^n$. Now, with every set of counts, there corresponds a set of frequencies

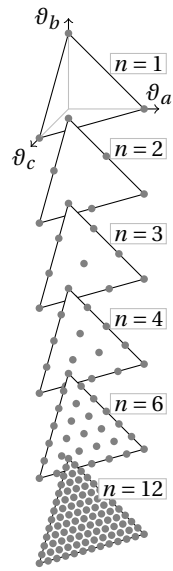
$$n^{\mathscr{X}}/n := \left\{ \tfrac{m}{n} \,\middle|\, m : n^{\mathscr{X}} \right\} \in \Delta_{\mathscr{X}} \cap \mathbb{Q}^{\mathscr{X}}, \tag{3.22}$$

where $\Delta_{\mathscr{X}}$ is the unit simplex for the sample space $\mathscr{X}$ (cf. $(1.56)_{51}$). We can define the coherent frequency distribution $\underline{R}^n$ on $\mathscr{L}_{n^{\mathscr{X}}/n}$ with

$$\underline{R}^n = \underline{Q}^n\big(\bullet \circ (\tfrac{1}{n}\cdot)\big) \qquad \text{or, equivalently,} \qquad \underline{Q}^n = \underline{R}^n\big(\bullet \circ (n\cdot)\big). \tag{3.23}$$

Now the idea is the following: instead of considering a sequence distribution $\underline{P}$ on (some sufficiently rich part of) $\mathscr{L}_{\mathbb{N}_{>0} \to \mathscr{X}}$, we consider a frequency distribution $\underline{R}$ on (some sufficiently rich part of) $\mathscr{L}_{\Delta_{\mathscr{X}}}$ to model the information available about X. Its possibility space, the set of all frequency vectors $\Delta_{\mathscr{X}}$, appeals intuitively and – because of its boundedness – technically. Just as the finitely exchangeable marginals $(\underline{P}^n \mid n : \mathbb{N}_{>0})$

Let $\mathscr{X} := \{a, b, c\}$; elements of $n^{\mathscr{X}}/n$ are denoted with a dot:



contained everything of practical interest in the sequence distribution $\underline{P}$, they, through the corresponding count distributions $(\underline{Q}^n \mid n : \mathbb{N}_{>0})$, also contain everything of practical interest in $\underline{R}$. The distributions in these two $\mathbb{N}_{>0}$-tuples should therefore be recoverable from $\underline{R}$ as induced previsions (cf. §1.3.1$_{52}$).

Conjuring up the relationships between gambles on $\mathscr{X}^n$ or on $n^{\mathscr{X}}$ and gambles on $\Delta_{\mathscr{X}}$ that characterize the relationship of $\underline{R}$ and its induced previsions, and verifying the consistency of these relationships takes up the rest of this subsection. As we expect of $\underline{R}$ that it behaves as the frequency distributions in $(\underline{R}^n \mid n : \mathbb{N}_{>0})$, we look for inspiration to the relationships between gambles for induced distributions for finite sequence lengths n and $N : \mathbb{N}_{\geq n}$.

Using (3.23)$_{\frown}$ and (3.20)–(3.21)$_{102–103}$, we know that for any gamble h on $n^{\mathscr{X}}$ we must have that

$$\underline{Q}^n h = \underline{Q}^N(S_n^N h) = \underline{R}^N\big(S_n^N h \circ (n\cdot)\big). \tag{3.24}$$

This must hold for N that are arbitrarily large; to see how $S_n^N h \circ (n\cdot)$ behaves in this limit, we write its value in some frequency vector ϑ in $\Delta_{\mathscr{X}} \cap \mathbb{Q}^{\mathscr{X}}$ and assume N to be such that $\vartheta \in N^{\mathscr{X}}/N$:

$$(S_n^N h)(N\cdot\vartheta) = \textstyle\sum_{m : n^{\mathscr{X}}} \frac{|[N\cdot\vartheta - m]|\cdot|[m]|}{|[N\cdot\vartheta]|} \cdot hm.$$

The fraction under the sum is the only thing that is affected by the magnitude of N; we rewrite it, first by applying (3.12)$_{100}$:

$$\frac{|[N\cdot\vartheta - m]|\cdot|[m]|}{|[N\cdot\vartheta]|} = \frac{\frac{(N-n)!}{\prod_{z:\mathscr{X}}(N\cdot\vartheta_z - m_z)!}\cdot|[m]|}{\frac{N!}{\prod_{z:\mathscr{X}}(N\cdot\vartheta_z)!}};$$

it should be understood that the right-hand side is zero whenever $\vartheta_z = 0$ and $m_z > 0$ for some z in $\mathscr{X}$, as then $[N\cdot\vartheta - m] = \emptyset$:

$$\begin{aligned} &= |[m]| \cdot \frac{\prod_{z:\mathscr{X}\wedge m_z>0}\prod_{\ell:0..m_z-1}(N\cdot\vartheta_z - \ell)}{\prod_{\ell:0..n-1}(N-\ell)} \\ &= |[m]| \cdot \textstyle\prod_{z:\mathscr{X}_m} \vartheta_z^{m_z} \cdot \dfrac{\prod_{\ell:1..m_z-1}(1-\frac{\ell}{N\cdot\vartheta_z})}{\prod_{\ell:1..n-1}(1-\frac{\ell}{N})}, \end{aligned}$$

where we have used a notational shorthand that will be useful later on too: for any α in $\mathbb{R}^{\mathscr{X}}$ such that $\alpha \geq 0$ we let $\mathscr{X}_\alpha := \{z : \mathscr{X} \mid \alpha_z > 0\}$. Finally,

$$\lim_{k\to+\infty} \tfrac{|[k!\cdot\vartheta - m]|\cdot|[m]|}{|[k!\cdot\vartheta]|} = B_m\vartheta := |[m]| \cdot \textstyle\prod_{z:\mathscr{X}_m} \vartheta_z^{m_z}, \tag{3.25}$$

and therefore

$$\lim_{k\to+\infty}(S_n^{k!}h)(k!\cdot\vartheta) = \mathrm{Cm}(h|n,\vartheta) := \textstyle\sum_{m:n^{\mathscr{X}}} hm\cdot B_m\vartheta, \tag{3.26}$$

where on the left-hand sides we consider the limiting sequence to start at index $\min\{k : \mathbb{N}_{>0} \mid \vartheta \in k!^{\mathscr{X}}/k!\}$.

The left-hand side of (3.25) provides – after extension to irrational frequency vectors by continuity – a definition of functions $B_m : \Delta_{\mathscr{X}} \to \mathbb{R}_{\geq 0}$, the (multivariate) Bernstein polynomials of degree n [Prautzsch et al. 2002, §10.1_{141}]. These form a basis for the polynomials on the unit simplex $\Delta_{\mathscr{X}}$ of degree up to n: for any N in $\mathbb{N}_{\geq n}$, any one of them, call it ν, is uniquely defined by some gamble b_ν^N on $n^{\mathscr{X}}$, which gives the coefficients of its decomposition in Bernstein polynomials, i.e., $\nu = \mathrm{Cm}(b_\nu^N | N, \bullet)$ (cf. (3.26)). The set of all polynomials on $\Delta_{\mathscr{X}}$ – a linear subspace of $\mathscr{L}_{\Delta_{\mathscr{X}}}$ – is denoted by

$$\mathscr{V}_{\Delta_{\mathscr{X}}} := \left\{ \mathrm{Cm}(h|n, \bullet) \mid n : \mathbb{N}_{>0} ; h : \mathscr{L}_{n^{\mathscr{X}}} \right\}. \tag{3.27}$$

In (3.26), we have introduced the linear prevision associated to what we call the count-multinomial distribution, which gives the probability of drawing a count vector m with sum n *with replacement* from an urn with a composition characterized by the frequency ϑ.

Based on the intuition gained by calculating the limits above, we modify (3.24) into

$$\underline{Q}^n h = \underline{R}\big(\mathrm{Cm}(h|n, \bullet)\big) \tag{3.28}$$

and make this our proposal for the relationship between $\underline{R}$ and $\underline{Q}^n$; it is a constraint that has to be satisfied for all n in $\mathbb{N}_{>0}$ and all gambles h on $n^{\mathscr{X}}$. It defines $\underline{R}$ uniquely on the set of all polynomial gambles $\mathscr{V}_{\Delta_{\mathscr{X}}}$. To show that this is a good proposal, we have to prove that this $\underline{R}$ is a *coherent* lower prevision on $\mathscr{V}_{\Delta_{\mathscr{X}}}$. On top of this, we must not forget that $(\underline{Q}^n \mid n : \mathbb{N}_{>0})$ is a tuple of count distributions for an exchangeable sequence of random variables X, which means that its components are linked by a consistency relationship of the form $\underline{Q}^N\big(S_n^N h\big) = \underline{Q}^n h$ (cf. $(3.20)_{102}$), for all N in $\mathbb{N}_{>n}$.

To verify that $\underline{R}$ respects this consistency relationship, let ν in $\mathscr{V}_{\Delta_{\mathscr{X}}}$ be of degree n, then $\underline{R}\nu = \underline{Q}^n(b_\nu^n)$, but also $\underline{R}\nu = \underline{Q}^N(b_\nu^N)$. This is consistent only when $\underline{Q}^N(b_\nu^N) = \underline{Q}^n(b_\nu^n)$, which is indeed guaranteed by $b_\nu^N = S_n^N b_\nu^n$ [Zhou's formula, Prautzsch et al. 2002, §11.9_{165}] and $(3.20)_{102}$.

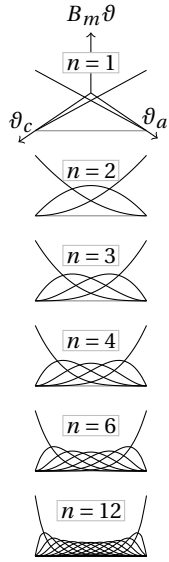
Now that we know that the values of $\underline{R}$ can be consistently assigned, we only need to check that this does not conflict with the coherence requirement. Given that $\mathscr{V}_{\Delta_{\mathscr{X}}}$ is a linear space, coherence is equivalent to superlinearity $(1.32)_{42}$ plus accepting sure gains $(1.29)_{42}$.

(i) Superlinearity: (let ν and w be polynomials of degree up to n and let λ and μ be nonnegative reals)

$$\begin{aligned} \underline{R}(\lambda \cdot \nu + \mu \cdot w) = \underline{Q}^n b^n_{\lambda\cdot\nu+\mu\cdot w} &= \underline{Q}^n(\lambda \cdot b_\nu^n + \mu \cdot b_w^n) \\ &\geq \lambda \cdot \underline{Q}^n b_\nu^n + \mu \cdot \underline{Q}^n b_w^n = \lambda \cdot \underline{R}\nu + \mu \cdot \underline{R} w, \end{aligned}$$

where $b^n_{\lambda\cdot\nu+\mu\cdot w} = \lambda \cdot b_\nu^n + \mu \cdot b_w^n$ because b_ν^n provides the coefficients of the *linear* decomposition of ν into Bernstein polynomials; the inequality follows from the coherence of $\underline{Q}^n$.

We plot the monomodal Bernstein polynomials of degree n (which sum to 1) on $\Delta_{\{a,c\}}$:



(ii) Accepting sure gains: (let v be a polynomial of degree n)

$$\underline{R}v \geq \sup_{N:\mathbb{N}_{\geq n}} \min\{b_v^N\} = \min\{v\};$$

the inequality follows from the fact that $\underline{R}v = \underline{Q}^N b_v^N$ for all N in $\mathbb{N}_{\geq n}$ and from the coherence of these $\underline{Q}^N$; the equality follows from the linear and thus uniform convergence of b_v^N to v for increasing N [Prautzsch et al. 2002, §11.9$_{165}$; Trump & Prautzsch 1996].

So now that we know we can find a frequency distribution $\underline{R}$, what is its use? It allows us to conveniently abstract away the essence of the assumption that the exchangeable sequence of samples we are dealing with is arbitrarily long: The frequency distribution $\underline{R}$ is a representation for all the information contained in the tuple $(\underline{Q}^n \mid n:\mathbb{N}_{>0})$ of count distributions apart from this assumption, i.e., that any pair of this tuple is consistent in the sense of (3.20)$_{102}$. This is formalized with the predicate $\infty\text{-cns}:\bigtimes_{n:\mathbb{N}_{>0}}(\mathscr{P}\mathscr{L}_{n^{\mathscr{X}}})_{\text{coh}} \to \mathbb{B}$ defined by

$$\begin{aligned}\infty\text{-cns}(\underline{Q}^n \mid n:\mathbb{N}_{>0}) \Leftrightarrow \forall n:\mathbb{N}_{>0};&\\ \underline{Q}^n = \underline{R}\big(\text{Cm}(\bullet|n,\bullet)\big),&\\ \text{with } \underline{R} := \underline{Q}^n b^n_{\bullet}.&\end{aligned} \tag{3.29}$$

Of course, we know that the count distributions already abstract away an implicit finite exchangeability assumption. The uncertainty model we started with was the tuple $(\underline{P}^n \mid n:\mathbb{N}_{>0})$ of marginal distributions for the vectors of random variables $(X_k \mid k:1..n)$. For them, we can abstract away the assumption that we are dealing with an arbitrarily long exchangeable sequence of samples: The frequency distribution $\underline{R}$ is a representation for all the information contained in the tuple $(\underline{P}^n \mid n:\mathbb{N}_{>0})$, apart from this assumption. This is formalized with the predicate $\infty\text{-xch}:\bigtimes_{n:\mathbb{N}_{>0}}(\mathscr{P}\mathscr{L}_{\mathscr{X}^n})_{\text{coh}} \to \mathbb{B}$ defined by

$$\begin{aligned}\infty\text{-xch}(\underline{P}^n \mid n:\mathbb{N}_{>0}) \Leftrightarrow \forall n:\mathbb{N}_{>0};&\\ \underline{P}^n = \underline{R}\big(\text{Mn}(\bullet|n,\bullet)\big),&\\ \text{with } \underline{R} := \underline{P}^n(b^n_{\bullet} \circ C_{\mathscr{X}}),&\end{aligned} \tag{3.30}$$

where, for every ϑ in $\Delta_{\mathscr{X}}$, $\text{Mn}(\bullet|n,\vartheta)$ is the linear prevision associated to the so-called multinomial distribution [Johnson et al. 1997, §35.2$_{31}$], which – not surprisingly – gives the probability of drawing a sequence of samples x of length n *with replacement* from an urn with a composition characterized by the frequency ϑ. It is related to the count-multinomial by the following relation, which also results in an explicit expression: (respectively use (3.19)$_{102}$, (3.26)$_{104}$, (3.18)$_{101}$ and (3.25)$_{104}$, and (3.9)$_{100}$)

$$\text{Mn}(f|n,\vartheta) := \text{Cm}\big(\text{Mh}(f|\bullet) \,\big|\, n,\vartheta\big) \tag{3.31}$$

$$= \textstyle\sum_{m:n^{\mathscr{X}}} \text{Mh}(f|m)\cdot B_m\vartheta \tag{3.32}$$

$$= \sum_{m:n^{\mathscr{X}}} \tfrac{1}{|[m]|} \cdot \sum_{y:[m]} f\,y \cdot |[m]| \cdot \prod_{z:\mathscr{X}_m} \vartheta_z^{m_z}$$
$$= \sum_{y:\mathscr{X}^n} f\,y \cdot B_y\vartheta, \tag{3.33}$$

where we have used a modified type of Bernstein polynomial defined for any sequence $y:\mathscr{X}^n$ by

$$B_y\vartheta := \prod_{z:\{y\}} \vartheta_z^{(C_{\mathscr{X}}y)_z} = \tfrac{1}{|[C_{\mathscr{X}}y]|} \cdot B_{C_{\mathscr{X}}y}\vartheta \tag{3.34}$$

for which it is useful to recall that $\{y\} = \mathscr{X}_{C_{\mathscr{X}}y} = \{z:\mathscr{X} \mid C_{\mathscr{X}}y > 0\}$.

The importance of this subsection's representation theorems (3.29) and (3.30) is that under an assumption of infinite exchangeability, we can – and perhaps should, to avoid not taking into account exchangeability – think in all generality in terms of frequency vectors (using a frequency distribution), any additional information given as supremum prices for gambles on $\mathscr{X}^n$ or on $n^{\mathscr{X}}$ can be translated to information for gambles in $\mathscr{V}_{\Delta_{\mathscr{X}}}$ using (3.29) or (3.30). These representation theorems are direct generalizations to coherent lower previsions of results by de Finetti [1974–1975, §11.4$_{\text{vol. 2, 215}}$] for linear previsions. Note, however, that in our derivation we did not need any intermediate step involving linear previsions. Also, our argument here is very different from any found in the literature and as such it constitutes a new proof of de Finetti's representation theorem.

A last remark before continuing with other matters: we know that any representing frequency distribution $\underline{R}$ is uniquely defined on $\mathscr{V}_{\Delta_{\mathscr{X}}}$, the set of all polynomials on $\Delta_{\mathscr{X}}$. But we can easily say more: the set of polynomials is uniformly dense in $\mathscr{C}_{\Delta_{\mathscr{X}}}$, the set of all continuous functions on $\Delta_{\mathscr{X}}$ [this is a consequence of the Stone-Weierstraß theorem, see, e.g., Pinkus 2005], which means $\underline{R}$ is also uniquely defined on $\mathscr{C}_{\Delta_{\mathscr{X}}}$ [Walley 1991, §2.6.1(l)$_{77}$]. It is not immediately clear whether $\underline{R}$ can be uniquely extended to even larger domains. However, what we do know is that there exists a – possibly nonunique – coherent extension to the whole of $\mathscr{L}_{\Delta_{\mathscr{X}}}$ [Walley 1991, §3.1.2$_{123}$].

3.1.4 *Posterior count distributions & sufficient statistics*

We now know how we should, in a situation of exchangeable sampling, represent the information we have about the still unobserved samples or the process that generates them: work with counts or frequencies. In this subsection, we investigate updating under finite exchangeability and show that updating the sequence distribution can be done entirely in terms of the corresponding count distribution and the count vector of the observed sample sequence. This investigation thereby shows that, once we have assumed exchangeability, the count vector of an observed sample sequence is a sufficient statistic, i.e., it contains all that is relevant for inference in that observed sample.

The following mnemonic may be useful: *Checked* variables (such as $\check{n}$) pertain to observed sequences, *hatted* ones (such as $\hat{n}$) to unobserved ones.

Consider the situation in which we have already observed $\check{n} : 1..N-1$ samples out of $N : \mathbb{N}_{>0}$ observations in total. The sample space is $\mathscr{X}$ and $\mathscr{X}^N$ is the corresponding possibility space; also let $\hat{n} := N - \check{n}$. Denote the sequence of observed samples by $\check{x} : \mathscr{X}^{\check{n}}$ and by $\check{m} := C_{\mathscr{X}}\check{x}$ the corresponding count vector. We also have some exchangeable set of desirable gambles $\mathscr{R} :\subset \mathscr{L}_{\mathscr{X}^N}$, the corresponding prior sequence distribution $\underline{P}$ on $\mathscr{L}_{\mathscr{X}^N}$, and its representing coherent prior count distribution $\underline{Q}$ on $\mathscr{L}_{N^{\mathscr{X}}}$ (cf. $(3.19)_{102}$).

The observed partial sequence of samples gives rise to three conditioning events, each of which differs only in the amount of information about the sequence order it preserves:

(i) $\imath\check{x} \times \mathscr{X}^{\hat{n}}$, which preserves all order information and which we can represent by $\check{x}$ itself;

(ii) $[\check{m}] \times \mathscr{X}^{\hat{n}}$, which discards the order in which the observed samples were encountered and which we can represent by $\check{m}$;

(iii) $(\mathscr{X}^N)_{\geq \check{m}} := \{y : \mathscr{X}^N \mid C_{\mathscr{X}} y \geq \check{m}\}$, which further discards how the observed samples were distributed among the *whole* sequence and thus preserves no order information at all; it can be identified with the count vector event $(N^{\mathscr{X}})_{\geq \check{m}}$; $\check{m}^+$ represents both events.

It is useful to have the expressions for the prevision of some gamble f on $\mathscr{X}^N$ respectively masked by each of these conditioning events. This involves the multivariate hypergeometric prevision $(3.18)_{101}$ of these masked gambles; we calculate these first: (let $m : (N^{\mathscr{X}})_{\geq \check{m}}$)

$$\begin{aligned}
\operatorname{Mh}\big(f \cdot I^{\imath\check{x}\times\mathscr{X}^{\hat{n}}} \,\big|\, m\big) &= \tfrac{1}{|[m]|} \cdot \textstyle\sum_{y:[m]} f\,y \cdot I^{\imath\check{x}\times\mathscr{X}^{\hat{n}}} y \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \tfrac{1}{|[m]|} \cdot \textstyle\sum_{\hat{y}:[m-\check{m}]} f(\check{x},\hat{y}) \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \tfrac{|[m-\check{m}]|}{|[m]|} \cdot \operatorname{Mh}\big(f(\check{x},\bullet) \,\big|\, m-\check{m}\big), \qquad (3.35)
\end{aligned}$$

$$\begin{aligned}
\operatorname{Mh}\big(f \cdot I^{[\check{m}]\times\mathscr{X}^{\hat{n}}} \,\big|\, m\big) &= \tfrac{1}{|[m]|} \cdot \textstyle\sum_{y:[m]} f\,y \cdot I^{[\check{m}]\times\mathscr{X}^{\hat{n}}} y \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \tfrac{1}{|[m]|} \cdot \textstyle\sum_{\hat{y}:[m-\check{m}]} \sum_{\check{y}:[\check{m}]} f(\check{y},\hat{y}) \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \tfrac{|[m-\check{m}]|\cdot|[\check{m}]|}{|[m]|} \cdot \operatorname{Mh}\big(\operatorname{Mh}(f_{[\check{m}]\times[m-\check{m}]}|\check{m}) \,\big|\, m-\check{m}\big), \qquad (3.36)
\end{aligned}$$

$$\begin{aligned}
\operatorname{Mh}\big(f \cdot I^{(\mathscr{X}^N)_{\geq\check{m}}} \,\big|\, m\big) &= \tfrac{1}{|[m]|} \cdot \textstyle\sum_{y:[m]} f\,y \cdot I^{(\mathscr{X}^N)_{\geq\check{m}}} y \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \tfrac{1}{|[m]|} \cdot \textstyle\sum_{y:m} f\,y \\
&= I^{(N^{\mathscr{X}})_{\geq\check{m}}} m \cdot \operatorname{Mh}(f|m). \qquad (3.37)
\end{aligned}$$

Each of the three resulting expressions is built up in the same fashion: the last factor uses the multivariate hypergeometric prevision to translate the masked gamble to a function on $N^{\mathscr{X}}$; the first factor is the translation of the mask to $N^{\mathscr{X}}$; any middle factors give an m-dependent weighting to this mask to compensate for the nonbijective character of

the mask translation process. The functions generating these compensating factors are respectively called sequence and count multivariate hypergeometric likelihood functions, which deserve their own symbol:

$$L_{\check{x}} := \mathrm{Mh}\big(\iota\check{x} \times \mathscr{X}^{\hat{n}} \,\big|\, \bullet\big) = m : N^{\mathscr{X}} \,;\, \tfrac{|[m - C_{\mathscr{X}}\check{x}]|}{|[m]|} \tag{3.38}$$

$$:= \hat{m} : \hat{n}^{\mathscr{X}} \,;\, \tfrac{|[\hat{m}]|}{|[C_{\mathscr{X}}\check{x} + \hat{m}]|}, \tag{3.39}$$

$$L_{\check{m}} := \mathrm{Mh}\big([\check{m}] \times \mathscr{X}^{\hat{n}} \,\big|\, \bullet\big) = m : N^{\mathscr{X}} \,;\, \tfrac{|[m - \check{m}]| \cdot |[\check{m}]|}{|[m]|} \tag{3.40}$$

$$:= \hat{m} : \hat{n}^{\mathscr{X}} \,;\, \tfrac{|[\hat{m}]| \cdot |[\check{m}]|}{|[\check{m} + \hat{m}]|}, \tag{3.41}$$

For both, we have given an additional definition that extends their domain to the space of count vectors for the unobserved samples. For each first definition, note that $L_{\check{x}} = I^{(N^{\mathscr{X}})_{\geq \check{m}}} \cdot L_{\check{x}}$ and $L_{\check{m}} = I^{(N^{\mathscr{X}})_{\geq \check{m}}} \cdot L_{\check{m}}$ because $|[m - \check{m}]| \propto I^{(N^{\mathscr{X}})_{\geq \check{m}}} m$ for all m in $N^{\mathscr{X}}$. Also note that $L_{\check{m}} = |[\check{m}]| \cdot L_{\check{x}}$ whenever $\check{x} \in [\check{m}]$.

Before moving on, let us give an illustrative annotated table that makes the quantities we have encountered so far more concrete. For this example, we take $\mathscr{X} := \{a, b\}$, $N := 4$, $\check{n} := 3$, and $\check{x} := (a, b, a)$ or aba; so then $\check{m} = (2, 1)$.

	m	$[m]$						$\lvert[m]\rvert$	$L_{\check{x}} m$	$L_{\check{m}} m$
	4,0	*aaaa*	$[\check{m}] \times \mathscr{X}^{\hat{n}}$					1	0	0
$(N^{\mathscr{X}})_{\geq \check{m}}$	3,1	*aaab*	*aaba*	*baaa*	*abaa*	$\iota\check{x} \times \mathscr{X}^{\hat{n}}$		4	1/4	3/4
	2,2	*bbaa*	*aabb*	*baab*	*abab*	*abba*	*baba*	6	1/6	3/6
	1,3	*bbba*	*bbab*	*abbb*	*babb*	$\{y : \mathscr{X}^N \mid C_{\mathscr{X}} y \geq \check{m}\}$		4	0	0
$N^{\mathscr{X}}$	0,4	*bbbb*					$\mathscr{X}^N$	1	0	0

The table makes it clear that the likelihood functions give – atom-by-atom – the relative number of sample sequences present in the event used. It is a useful exercise to choose some gamble on $\mathscr{X}^N$ (i.e., attach a value to each sequence) and see how it gets transformed to a gamble on $N^{\mathscr{X}}$; a constant gamble is already quite illuminating.

It is also a useful exercise to fix $\underline{P}$ – e.g., let it be uniform or vacuous – and see how the prevision of some gamble on $\mathscr{X}^N$ is determined by its corresponding gamble on $N^{\mathscr{X}}$. To do this in general, combine the representation theorem (3.19)$_{102}$, equations (3.35)–(3.37) and definitions (3.38) and (3.40):

$$\underline{P}\big(f \cdot I^{\iota\check{x} \times \mathscr{X}^{\hat{n}}}\big) = \underline{Q}\Big(L_{\check{x}} \cdot \mathrm{Mh}\big(f(\check{x}, \bullet) \,\big|\, \bullet - \check{m}\big)\Big), \tag{3.42}$$

$$\underline{P}\big(f \cdot I^{[\check{m}] \times \mathscr{X}^{\hat{n}}}\big) = \underline{Q}\Big(L_{\check{m}} \cdot \mathrm{Mh}\big(\mathrm{Mh}(f_{[\check{m}] \times [\bullet - \check{m}]} \,|\, \check{m}) \,\big|\, \bullet - \check{m}\big)\Big), \tag{3.43}$$

$$\underline{P}\big(f \cdot I^{(\mathscr{X}^N)_{\geq \check{m}}}\big) = \underline{Q}\big(I^{(N^{\mathscr{X}})_{\geq \check{m}}} \cdot \mathrm{Mh}(f \,|\, \bullet)\big). \tag{3.44}$$

Now, when we let f be identically one, we get the lower probabilities of the conditioning events:

$$\begin{aligned}\underline{P}(\imath\check{x}\times\mathscr{X}^{\hat{n}}) &= \underline{Q}L_{\check{x}} && (3.45)\\ &\geq \underline{Q}(N^{\mathscr{X}})_{\geq\check{m}}\cdot\overbrace{\min_{\hat{m}:\hat{n}^{\mathscr{X}}}L_{\check{x}}\hat{m}}^{>0},\\ \underline{P}([\check{m}]\times\mathscr{X}^{\hat{n}}) &= \underline{Q}L_{\check{m}} && (3.46)\\ &\geq \underline{Q}(N^{\mathscr{X}})_{\geq\check{m}}\cdot\overbrace{\min_{\hat{m}:\hat{n}^{\mathscr{X}}}L_{\check{m}}\hat{m}}^{>0},\\ \underline{P}(\mathscr{X}^{N})_{\geq\check{m}} &= \underline{Q}(N^{\mathscr{X}})_{\geq\check{m}} && (3.47)\\ &\geq \underline{Q}L_{\check{m}} = |[\check{m}]|\cdot\underline{Q}L_{\check{x}}.\end{aligned}$$

The included lower bounds allow us to conclude that these lower probabilities are either all zero or all strictly positive (in which case they are listed in order of increasing magnitude). So we can use the GBR (1.82)–$(1.83)_{61}$ to obtain the updated previsions for either none or all three of the events. Completely analogous results for the corresponding upper probabilities imply that regular extension $(1.70)_{58}$ can make a difference either for all or none of the three events.

At this point we have – in (3.42)–$(3.47)_{109–110}$ – all the elements necessary to obtain the exchangeable updated sequence distributions

(i) $\underline{P}(\bullet|\check{x})$ on $\mathscr{L}_{\mathscr{X}^{\hat{n}}}$, where we have identified $\imath\check{x}\times\mathscr{X}^{\hat{n}}$ with $\mathscr{X}^{\hat{n}}$,
(ii) $\underline{P}(\bullet|\check{m})$ on $\mathscr{L}_{[\check{m}]\times\mathscr{X}^{\hat{n}}}$, and
(iii) $\underline{P}(\bullet|\check{m}^{+})$ on $\mathscr{L}_{(\mathscr{X}^{N})_{\geq\check{m}}}$.

These are exchangeable in the sense that they correspond to updated sets of desirable gambles $\mathscr{R}_{\imath\check{x}\times\mathscr{X}^{\hat{n}}}$, $\mathscr{R}_{[\check{m}]\times\mathscr{X}^{\hat{n}}}$, and $\mathscr{R}_{(\mathscr{X}^{N})_{\geq\check{m}}}$ that satisfy a constraint such as $(3.5)_{97}$: (let A be $\imath\check{x}\times\mathscr{X}^{\hat{n}}$, $[\check{m}]\times\mathscr{X}^{\hat{n}}$, or $(\mathscr{X}^{N})_{\geq\check{m}}$)

$$\mathscr{H}A+(\mathscr{R}_{A})_{\neq(A;0)}\subseteq\mathscr{R}_{A}, \tag{3.48}$$

where

$$\mathscr{H}A:=\operatorname{span}\{\pi g-g|g:\mathscr{L}_{A};\pi:\Pi_{1..N}\wedge\pi A=A\} \tag{3.49}$$

is the linear subspace of gambles in $\mathscr{L}_{A}$ that are marginally desirable because of a conditional exchangeability assumption.

Constraint (3.48) holds, because updating sets of desirable gambles preserves exchangeability: conditioning both sides of $(3.5)_{97}$ on A gives

$$\begin{aligned}\mathscr{R}_{A}&\supseteq(\mathscr{H}\mathscr{X}^{N}+\mathscr{R}_{\neq(\mathscr{X}^{N};0)})_{A}\\ &\supseteq(\mathscr{H}\mathscr{X}^{N})_{A}+(\mathscr{R}_{\neq(\mathscr{X}^{N};0)})_{A}\supseteq\mathscr{H}A+(\mathscr{R}_{A})_{\neq(A;0)},\end{aligned}$$

where the last step follows from $(\mathscr{R}_{A})_{\neq(A;0)}\subseteq(\mathscr{R}_{\neq(\mathscr{X}^{N};0)})_{A}$ and

$$\begin{aligned}\mathscr{H}A&=\operatorname{span}\{(\pi f-f)_{A}|f:\mathscr{L}_{\mathscr{X}^{N}}\wedge\operatorname{supp}f\subseteq A;\pi:\Pi_{1..N}\wedge\pi A=A\}\\ &\subseteq\operatorname{span}\{(\pi f-f)_{A}|f:\mathscr{L}_{\mathscr{X}^{N}};\pi:\Pi_{1..N}\}\\ &\subseteq(\mathscr{H}\mathscr{X}^{N})_{A}.\end{aligned}$$

This constraint gives rise to representation theorems like $(3.19)_{102}$: with $\underline{P}(\bullet|\check{x})$, $\underline{P}(\bullet|\check{m})$, and $\underline{P}(\bullet|\check{m}^+)$ there correspond coherent count distributions $\underline{Q}(\bullet|\check{x})$, $\underline{Q}(\bullet|\check{m})$, and $\underline{Q}(\bullet|\check{m}^+)$ on $\mathscr{L}_{\hat{n}^{\mathscr{X}}}$ such that (let g, h, and f be gambles on $\mathscr{X}^{\hat{n}}$, $[\check{m}]\times\mathscr{X}^{\hat{\bar{n}}}$, and $(\mathscr{X}^N)_{\geq\check{m}}$, respectively)

$$\underline{P}(g|\check{x}) = \underline{Q}\big(\mathrm{Mh}(g|\bullet)\,\big|\,\check{x}\big), \tag{3.50}$$

$$\underline{P}(h|\check{m}) = \underline{Q}\Big(\mathrm{Mh}\big(\mathrm{Mh}(h_{[\check{m}]\times[\bullet]}|\check{m})\,\big|\,\bullet\big)\,\Big|\,\check{m}\Big), \tag{3.51}$$

$$\underline{P}(f|\check{m}^+) = \underline{Q}\big(\mathrm{Mh}(f|\check{m}+\bullet)\,\big|\,\check{m}^+\big). \tag{3.52}$$

These representations follow from a completely analogous reasoning as for $(3.19)_{102}$, but whereas in $(3.16)_{101}$ and $(3.17)_{101}$ the gamble transformation of interest was $\frac{1}{N!}\cdot\sum_{\pi:\Pi_{1..N}}\pi$, it now is $\frac{1}{|\{\pi:\Pi_{1..N}|\pi A=A\}|}\cdot\sum_{\pi:\Pi_{1..N}\wedge\pi A=A}\pi$. The conceptually most important steps for working this out in full can be found in (3.35)–$(3.37)_{108}$.

The count distributions can be defined via natural extension $(1.83)_{61}$ of $\underline{P}$ to $\underline{P}(\bullet|\check{x})$, $\underline{P}(\bullet|\check{m})$, and $\underline{P}(\bullet|\check{m}^+)$, respectively; furthermore, $(3.6)_{99}$ is used to go from $\mathscr{M}\underline{P}$ to $\mathscr{M}\underline{Q}$: (this time, let $f:\mathscr{L}_{\hat{n}^{\mathscr{X}}}$)

As $\underline{P}$ and $\underline{Q}$ are assumed to be coherent here, $\mathrm{lce}_{\underline{P}}=\underline{P}$ and $\mathrm{lce}_{\underline{Q}}=\underline{Q}$.

$$\underline{Q}(f|\check{x}) = \begin{cases}\min_{Q:\mathrm{ext}(\mathscr{M}\underline{Q})}\frac{1}{Q[L_{\check{x}}}\cdot Q(L_{\check{x}}\cdot f_{N^{\mathscr{X}}}), & \underline{Q}(N^{\mathscr{X}})_{\geq C_{\mathscr{X}}\check{x}}>0,\\ \min\{f\}, & \text{otherwise;}\end{cases} \tag{3.53}$$

$$\underline{Q}(f|\check{m}) = \begin{cases}\min_{Q:\mathrm{ext}(\mathscr{M}\underline{Q})}\frac{1}{Q[L_{\check{m}}}\cdot Q(L_{\check{m}}\cdot f_{N^{\mathscr{X}}}), & \underline{Q}(N^{\mathscr{X}})_{\geq\check{m}}>0,\\ \min\{f\}, & \text{otherwise;}\end{cases} \tag{3.54}$$

$$\underline{Q}(f|\check{m}^+) = \begin{cases}\min_{Q:\mathrm{ext}(\mathscr{M}\underline{Q})}\frac{1}{Q(N^{\mathscr{X}})_{\geq\check{m}}}\cdot Qf_{N^{\mathscr{X}}}, & \underline{Q}(N^{\mathscr{X}})_{\geq\check{m}}>0,\\ \min\{f\}, & \text{otherwise.}\end{cases} \tag{3.55}$$

To get the expressions for updating using regular extension (cf. $(1.70)_{58}$), replace the condition '$\underline{Q}(N^{\mathscr{X}})_{\geq\bullet}>0$' by '$\overline{Q}(N^{\mathscr{X}})_{\geq\bullet}>0$' and the minimum by an infimum over $\{Q:\mathscr{M}\underline{Q}\,|\,Q(N^{\mathscr{X}})_{\geq\bullet}>0\}$.

There are two major final remarks to make about these updated sequence distributions and their corresponding count distribution:

(i) Observe that $\underline{Q}(\bullet|\check{x})=\underline{Q}(\bullet|C_{\mathscr{X}}\check{x})$, as $L_{C_{\mathscr{X}}\check{x}}=|[C_{\mathscr{X}}\check{x}]|\cdot L_{\check{x}}$: let the gamble h used in (3.51) be the cylindrical extension (cf. §1.3.1$_{52}$) to $[C_{\mathscr{X}}\check{x}]\times\mathscr{X}^{\hat{n}}$ of the gamble g used in (3.50), then

$$\underline{P}(g|\check{x}) = \underline{Q}\big(\mathrm{Mh}(g|\bullet)\,\big|\,\check{x}\big) = \underline{Q}\big(\mathrm{Mh}(h|\bullet)\,\big|\,C_{\mathscr{X}}\check{x}\big) = \underline{P}(h|C_{\mathscr{X}}\check{x}).$$

So we see that, for the count distributions, we can drop one of the two notations: we choose to drop the first, as the second clearly shows that the order of $\check{x}$ is irrelevant in the updating process (under an assumption of exchangeability). The sample sequence order is therefore called an ancillary statistic, in contrast to the name sufficient statistic that is given to the count vector.

(ii) From (3.55), we see that the prevision $\underline{Q}(\bullet|\check{m}^+)$ is both the count distribution corresponding to $\underline{P}(\bullet|\check{m}^+)$ as well as the updated lower

prevision obtained by conditioning $\underline{Q}$ on $(N^{\mathscr{X}})_{\geq \check{m}}$ – up to the identification of $(N^{\mathscr{X}})_{\geq \check{m}}$ with $\hat{n}^{\mathscr{X}}$. The count distribution $\underline{Q}(\bullet|\check{m})$ (and thus $\underline{Q}(\bullet|\check{x})$) is not an updated lower prevision of $\underline{Q}$, even though it is defined on a possibility space $\hat{n}^{\mathscr{X}}$ that is isomorphic to the subset $(N^{\mathscr{X}})_{\geq \check{m}}$ of the possibility space $N^{\mathscr{X}}$ of $\underline{Q}$. To terminologically position it in relation to the prior count distribution $\underline{Q}$ it is derived from, it is called a posterior count distribution.

A closing reminder: whenever natural and regular extension differ, or when they are both vacuous, other previsions can be jointly coherent updated previsions (cf. the last two paragraphs before §1.3.4$_{59}$).

3.1.5 *Posterior frequency distributions & sufficient statistics*

In the last subsection, we have gotten our feet wet with updating under the *finite* exchangeability assumption. In this subsection, we are going to dive headfirst into the foam-headed waves by updating under the *infinite* exchangeability assumption of §3.1.3$_{103}$. We show that updating a marginal sequence or count distribution can be done entirely in terms of the corresponding frequency distribution and the count vector of the observed sample. Our investigation thereby again shows that, once we have assumed exchangeability, the count vector of an observed sample is a sufficient statistic, i.e., it contains all that is relevant for inference in that observed sample.

We now consider the following setting: we have observed $\check{n}:\mathbb{N}_{>0}$ samples out of an infinite number of total possible observations. The sample space is $\mathscr{X}$ and $\mathbb{N}_{>0} \to \mathscr{X}$ is the corresponding possibility space. Denote the partial sequence of samples by $\check{x}:\mathscr{X}^{\check{n}}$ and by $\check{m} := C_{\mathscr{X}}\check{x}$ the corresponding count vector. We also have some coherent prior frequency distribution $\underline{R}$ on $\mathscr{C}_{\Delta_{\mathscr{X}}}$ and are going to look at what we can learn from updating the corresponding exchangeable marginal sequence distribution $\underline{P}^n$ on $\mathscr{L}_{\mathscr{X}^n}$ and coherent count distribution $\underline{Q}^n$ on $\mathscr{L}_{n^{\mathscr{X}}}$, where n is some number in $\mathbb{N}_{>\check{n}}$ (cf. (3.29)$_{106}$ and (3.30)$_{106}$). Furthermore, we introduce the shorthand $\hat{n} := n - \check{n}$.

Contrary to §3.1.4$_{107}$, we here only investigate two of the three conditioning events for $\underline{P}^n$ the partial sequence of samples gives rise to: $\iota\check{x} \times \mathscr{X}^{\hat{n}}$ and $[\check{m}] \times \mathscr{X}^{\hat{n}}$. These events are represented by $\check{x}$ and $\check{m}$. The third, $(\mathscr{X}^n)_{\geq \check{m}} := \{y:\mathscr{X}^n \mid C_{\mathscr{X}}y \geq \check{m}\}$ would lead to posteriors that are of little interest to us; although $(n^{\mathscr{X}})_{\geq \check{m}}$ the one conditioning event for $\underline{Q}^n$ does play a supporting role.

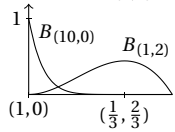
It is useful to have the expression for the prevision of some gamble f on $\mathscr{X}^n$ or h on $n^{\mathscr{X}}$, masked by the conditioning events above. This involves the multinomial prevision (3.33)$_{107}$ and count-multinomial prevision (3.26)$_{104}$ of these masked gambles; we start with h: (let $\vartheta:\Delta_{\mathscr{X}}$)

$$\mathrm{Cm}\big(h \cdot I^{(n^{\mathscr{X}})_{\geq \check{m}}} \,\big|\, n, \vartheta\big) = \textstyle\sum_{m:n^{\mathscr{X}}} hm \cdot I^{(n^{\mathscr{X}})_{\geq \check{m}}} m \cdot B_m\vartheta$$

$$= \sum_{m:(n^{\mathscr{X}})_{\geq \check{m}}} hm \cdot \tfrac{B_{\check{m}}\vartheta \cdot B_{m-\check{m}}\vartheta}{L_{\check{m}}m}$$
$$= B_{\check{m}}\vartheta \cdot \sum_{\hat{m}:\hat{n}^{\mathscr{X}}} \tfrac{1}{L_{\check{m}}\hat{m}} \cdot h(\check{m}+\hat{m}) \cdot B_{\hat{m}}\vartheta$$
$$= B_{\check{m}}\vartheta \cdot \mathrm{Cm}\big(\tfrac{h(\check{m}+\bullet)}{L_{\check{m}}} \,\big|\, \hat{n}, \vartheta\big), \tag{3.56}$$

where for the second equality, the identity $L_{\check{m}}m \cdot B_m = B_{\check{m}} \cdot B_{m-\check{m}}$ was used, which follows from $(3.25)_{104}$ and $(3.40)_{109}$; the third equality involves $(3.41)_{109}$. This expression is built up as follows: the last factor uses the count-multinomial prevision to translate the masked gamble to a function on the unit simplex $\Delta_{\mathscr{X}}$; the mask itself is translated to the whole of the unit simplex; the first factor gives a ϑ-dependent and $\check{m}$-specific weighting over the unit simplex. The Bernstein polynomial

Two count-multinomial likelihoods on $\Delta_{\{a,b\}}$:



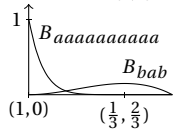
$$B_{\check{m}} = \mathrm{Cm}\big(L_{\check{m}} \,\big|\, \hat{n}, \bullet\big)$$
$$= \mathrm{Cm}\Big(\mathrm{Mh}\big([\check{m}] \times \mathscr{X}^{\hat{n}} \,\big|\, \bullet\big) \,\Big|\, \hat{n}, \bullet\Big) = \mathrm{Mn}\big([\check{m}] \times \mathscr{X}^{\hat{n}} \,\big|\, \hat{n}, \bullet\big) \tag{3.57}$$

assigning these weights can therefore be seen as a count-multinomial likelihood function (we used $(3.40)_{109}$ and $(3.31)_{106}$). But let us return to our original goal: expression (3.56) can be used together with $(3.31)_{106}$, (3.35)–$(3.36)_{108}$, and (3.38)–$(3.41)_{109}$ to obtain the multinomial prevision of the masked versions of f:

$$\mathrm{Mn}\big(f \cdot I^{\iota\check{x}\times\mathscr{X}^{\hat{n}}} \,\big|\, n, \vartheta\big) = B_{\check{x}}\vartheta \cdot \mathrm{Mn}\big(f(\check{x}, \bullet) \,\big|\, \hat{n}, \vartheta\big), \tag{3.58}$$
$$\mathrm{Mn}\big(f \cdot I^{[\check{m}]\times\mathscr{X}^{\hat{n}}} \,\big|\, n, \vartheta\big) = B_{\check{m}}\vartheta \cdot \mathrm{Mn}\big(\mathrm{Mh}(f_{[\check{m}]\times\mathscr{X}^{\hat{n}}}|\check{m}) \,\big|\, \hat{n}, \vartheta\big), \tag{3.59}$$

where in (3.58) we have used a modified Bernstein polynomial $(3.34)_{107}$, which can be seen as a sequence multinomial likelihood function

Two sequence multinomial likelihoods on $\Delta_{\{a,b\}}$:



$$B_{\check{x}} = \mathrm{Mn}\big(\iota\check{x} \times \mathscr{X}^{\hat{n}} \,\big|\, \hat{n}, \bullet\big). \tag{3.60}$$

Note that its relationship with $B_{\check{m}}$ is in complete analogy to the relationship between $L_{\check{x}}$ and $L_{\check{m}}$ observed in the previous subsection (cf. expressions (3.38)–$(3.41)_{109}$ and below).

We can now combine the representation theorem $(3.30)_{106}$, equations (3.56)–(3.59), and definition $(3.34)_{107}$ to obtain

$$\underline{P}^n\big(f \cdot I^{\iota\check{x}\times\mathscr{X}^{\hat{n}}}\big) = \underline{R}\Big(B_{\check{x}} \cdot \mathrm{Mn}\big(f(\check{x}, \bullet) \,\big|\, \hat{n}, \bullet\big)\Big), \tag{3.61}$$
$$\underline{P}^n\big(f \cdot I^{[\check{m}]\times\mathscr{X}^{\hat{n}}}\big) = \underline{R}\Big(B_{\check{m}} \cdot \mathrm{Mn}\big(\mathrm{Mh}(f_{[\check{m}]\times\mathscr{X}^{\hat{n}}}|\check{m}) \,\big|\, \hat{n}, \bullet\big)\Big). \tag{3.62}$$

When we let h and f be identically one, we get the lower probabilities of the conditioning events:

$$\underline{P}^n(\iota\check{x} \times \mathscr{X}^{\hat{n}}) = \underline{R}B_{\check{x}} = \underline{R}B_{C_{\mathscr{X}}\check{x}} \cdot \tfrac{1}{|[C_{\mathscr{X}}\check{x}]|}, \tag{3.63}$$
$$\underline{P}^n\big([\check{m}] \times \mathscr{X}^{\hat{n}}\big) = \underline{R}B_{\check{m}}. \tag{3.64}$$

Whenever $C_{\mathscr{X}}\check{x} = \check{m}$, both lower probabilities are either zero or strictly positive (in which case they are listed in order of increasing magnitude).

So we can use the GBR (1.82)–$(1.83)_{61}$ to obtain the updated previsions for either none or both of the events. Completely analogous results for the corresponding upper probabilities imply that regular extension $(1.70)_{58}$ can make a difference either for both or none of the events.

At this point we have – in (3.61)–$(3.64)_{\frown}$ – all the elements necessary to obtain, for all $n > \check{n}$ and thus for all $\hat{n} > 0$, the updated exchangeable marginal sequence distributions and via (3.50)–$(3.51)_{111}$ and (3.53)–$(3.54)_{111}$ their induced count distributions:

(i) $\underline{P}^{\hat{n}}(\bullet|\check{x})$ on $\mathscr{L}_{\mathscr{X}^{\hat{n}}}$ and $\underline{Q}^{\hat{n}}(\bullet|\check{x})$ on $\mathscr{L}_{\hat{n}^{\mathscr{X}}}$ and

(ii) $\underline{P}^{\hat{n}}(\bullet|\check{m})$ on $\mathscr{L}_{[\check{m}]\times\mathscr{X}^{\hat{n}}}$ and $\underline{Q}^{\hat{n}}(\bullet|\check{m})$ on $\mathscr{L}_{\hat{n}^{\mathscr{X}}}$.

We introduce two coherent lower previsions on $\mathscr{C}_{\Delta_{\mathscr{X}}}$, respectively $\underline{R}(\bullet|\check{x})$ and $\underline{R}(\bullet|\check{m})$, that are such that ($g$ and h are gambles on $\mathscr{X}^{\hat{n}}$ and $[\check{m}]\times\mathscr{X}^{\hat{n}}$, respectively)

$$\underline{P}^{\hat{n}}(g|\check{x}) = \underline{R}\big(\mathrm{Mn}(g|\hat{n},\bullet)\,\big|\,\check{x}\big), \tag{3.65}$$

$$\underline{P}^{\hat{n}}(h|\check{m}) = \underline{R}\Big(\mathrm{Mn}\big(\mathrm{Mh}(h|\check{m})\,\big|\,\hat{n},\bullet\big)\,\Big|\,\check{m}\Big). \tag{3.66}$$

(now let $h : \mathscr{L}_{\hat{n}^{\mathscr{X}}}$)

$$\underline{Q}^{\hat{n}}(h|C_{\mathscr{X}}\check{x}) = \underline{R}\big(\mathrm{Cm}(h|\hat{n},\bullet)\,\big|\,\check{x}\big), \tag{3.67}$$

$$\underline{Q}^{\hat{n}}(h|\check{m}) = \underline{R}\big(\mathrm{Cm}(h|\hat{n},\bullet)\,\big|\,\check{m}\big). \tag{3.68}$$

These frequency distributions $\underline{R}(\bullet|\check{x})$ and $\underline{R}(\bullet|\check{m})$ are defined via natural extension $(1.83)_{61}$ of $\underline{P}^n$ or $\underline{Q}^n$, the representation theorems $(3.29)_{106}$ and $(3.30)_{106}$, and under the assumption that the part of an infinitely exchangeable sequence of random variables that remains after a partial observation is still infinitely exchangeable: (now let f be a continuous gamble on the compact unit simplex $\Delta_{\mathscr{X}}$)

$$\underline{R}(f|\check{x}) = \begin{cases} \min_{R:\mathrm{ext}(\mathscr{M}\underline{R})} \frac{1}{RB_{\check{x}}} \cdot R(B_{\check{x}} \cdot f), & \underline{R}B_{C_{\mathscr{X}}\check{x}} > 0, \\ \min\{f\}, & \text{otherwise;} \end{cases} \tag{3.69}$$

$$\underline{R}(f|\check{m}) = \begin{cases} \min_{R:\mathrm{ext}(\mathscr{M}\underline{R})} \frac{1}{RB_{\check{m}}} \cdot R(B_{\check{m}} \cdot f), & \underline{R}B_{\check{m}} > 0, \\ \min\{f\}, & \text{otherwise.} \end{cases} \tag{3.70}$$

To get the expressions for updating using regular extension (cf. $(1.70)_{58}$), replace the condition '$\underline{R}B_\bullet > 0$' by '$\overline{R}B_\bullet > 0$' and the minimum by an infimum over $\{R : \mathscr{M}\underline{R} \mid RB_\bullet > 0\}$.

There are again two major final remarks to make about these updated distributions and their corresponding frequency distribution:

(i) From the equations (3.65)–(3.68) above, we can see that

$$\infty\text{-cns}\big(\underline{Q}^{\hat{n}}(\bullet|C_{\mathscr{X}}\check{x})\,\big|\,\hat{n}:\mathbb{N}_{>0}\big), \qquad \infty\text{-xch}\big(\underline{P}^{\hat{n}}(\bullet|\check{x})\,\big|\,\hat{n}:\mathbb{N}_{>0}\big),$$
$$\infty\text{-cns}\big(\underline{Q}^{\hat{n}}(\bullet|\check{m})\,\big|\,\hat{n}:\mathbb{N}_{>0}\big), \qquad \infty\text{-xch}\big(\underline{P}^{\hat{n}}(\bullet|\check{m})\,\big|\,\hat{n}:\mathbb{N}_{>0}\big),$$

with representing frequency distributions $\underline{R}(\bullet|\check{x})$ and $\underline{R}(\bullet|\check{m})$ defined by (3.69)–(3.70) that moreover coincide. So again we see that,

for these frequency distributions, we can drop one of the two notations: just as for the count distributions, we choose to drop the first, as the second clearly shows that the order of $\check{x}$ is irrelevant in the updating process. So also here the sample sequence order is an ancillary statistic and the count vector is a sufficient statistic.

(ii) Of course $\underline{R}(\bullet|\check{m})$ (and thus $\underline{R}(\bullet|\check{x})$) is not an updated lower prevision of $\underline{R}$, despite the suggestive notation. This is already evident from the fact that it is defined on the same domain as $\underline{R}$, and thus on the same possibility space $\Delta_{\mathscr{X}}$. To terminologically position it in relation to the prior frequency distribution $\underline{R}$ they are derived from, it is called a posterior frequency distribution.

3.1.6 *Classical Bayesian updating, likelihood functions & predictive versus parametric inference*

It is instructive to relate the work we did in the last two subsections to classical Bayesian updating [Bernardo & Smith 1994, §5.1$_{241}$]. It allows us to place what we have done in a wider context.

In §3.1.4$_{107}$ and §3.1.5$_{112}$ likelihood functions appeared naturally in our derivation of updated finitely and infinitely exchangeable distributions and posterior count and frequency distributions. It can be educating to see how likelihood functions come into being, framed more generally and without the noise generated by other considerations.

Consider a finite possibility space Ω and assume that the uncertainty about which element will be observed is described by a linear prevision $P(\bullet|\xi)$ on $\mathscr{L}_\Omega$, the sampling model, where ξ is a parameter whose possible values form the finite set Ξ. Now assume that there is uncertainty about the sampling model's parameter ξ and model this uncertainty by a coherent lower prevision $\underline{Q}$ on $\mathscr{L}_\Xi$ which is called the (parametric) prior.

In such a context, it is useful to patch together the sampling models for all possible parameters into a conditional linear prevision $\underline{P}(\bullet|\bullet)$ on $\mathscr{L}_\Omega$, as it allows us to find the joint uncertainty model $\underline{E}$ on $\mathscr{L}_{\Omega\times\Xi}$ defined for any gamble f on $\Omega\times\Xi$ by $\underline{E}f := \underline{Q}\big(P(f|\bullet)\big)$ using the marginal extension theorem (1.84)$_{62}$. Its Ω-marginal, the coherent lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$ defined for all gambles h on Ω by $\underline{P}h := \underline{E}(h\cdot I^{\Xi}) = \underline{Q}\big(P(h|\bullet)\big)$ is called the predictive prior.

Now assume that at some point $A\subset\Omega$ is observed. As we equate updating with conditioning, we define the updated joint coherent lower prevision $\underline{E}(\bullet|A)$ on $\mathscr{L}_{A\times\Xi}$ using the GBR (1.83)$_{61}$: (let f be a gamble on $A\times\Xi$)

$$\underline{E}(f|A) = \begin{cases} \min_{E:\mathrm{ext}(\mathscr{M}\underline{E})} \frac{1}{E(A\times\Xi)}\cdot E(f_{\Omega\times\Xi}\cdot I^{A\times\Xi}), & \underline{E}(A\times\Xi)>0,\\ \min\{f\}, & \text{otherwise,}\end{cases} \tag{3.71}$$

where

$$E(f_{\Omega\times\Xi}\cdot I^{A\times\Xi}) = \underline{Q}\big(P(f_{\Omega\times\Xi}\cdot I^{A}|\bullet)\big).$$

When restricting attention to $f := g\cdot I^A$, with g a gamble on Ξ, we obtain the defining expression for the parametric posterior, the coherent lower prevision $\underline{Q}(\bullet|A)$ on $\mathscr{L}_\Xi$:

$$\underline{Q}(g|A) = \begin{cases} \min_{Q:\mathrm{ext}(\mathscr{M}\underline{Q})} \frac{1}{QK_A}\cdot Q\big(g\cdot K_A\big), & \underline{Q}K_A > 0, \\ \min\{g\}, & \text{otherwise}, \end{cases} \tag{3.72}$$

where $K_A := P(A|\bullet)$ is the likelihood function of the observation A [also see Walley 1991, §8.4.8$_{429}$]: it gives the probability of A as a function of the sampling model's parameter. When restricting attention to $f := h\cdot I^{\Xi}$, with h a gamble on A, we obtain the defining expression for the predictive posterior, the coherent lower prevision $\underline{P}(\bullet|A)$ on $\mathscr{L}_A$:

$$\underline{P}(h|A) = \begin{cases} \min_{Q:\mathrm{ext}(\mathscr{M}\underline{Q})} \frac{1}{QK_A}\cdot Q\big(P(h_\Omega\cdot I^A|\bullet)\big), & \underline{Q}K_A > 0, \\ \min\{h\}, & \text{otherwise}. \end{cases} \tag{3.73}$$

In many interesting cases, Ω, $P(\bullet|\bullet)$, and A are such that $P(h_\Omega\cdot I^A|\bullet)$ factorizes into $P(h_\Omega|\bullet)\cdot K_A$, and then $\underline{P}(h|A) = \underline{Q}\big(P(h_\Omega|\bullet)\,\big|\,A\big)$ for $\underline{Q}K_A > 0$. We call A the posterior possibility space.

To update, we could also have used regular extension (1.70)$_{58}$; replace the minimum in (3.72) and (3.73) over $\mathscr{M}\underline{Q}$ by an infimum over all Q in $\mathscr{M}\underline{Q}$ such that $Q\big(P(A|\varphi^{-1}\bullet)\big) > 0$ and replace the condition $\underline{Q}K_A > 0$ by $\overline{Q}K_A > 0$ [also see Walley 1991, §J5$_{640}$].

The likelihood principle can break down if the sampling model that lies at its basis is imprecise [Walley 1991, §8.6.9$_{440}$].

Notice that parametric posterior only depends on the observation through the *normalized* likelihood function. This observation is called the (finite) likelihood principle see, e.g., Bernardo & Smith [1994, §5.1.4$_{249}$].

Good illustrations of what we have just been discussing can be found in §3.1.2$_{99}$ and §3.1.3$_{103}$:

(i) We encountered two related assumptions: finite and infinite exchangeability, respectively leading to drawing without and with replacement from an urn as sampling models. The resulting parametric posteriors are defined by (3.53)–(3.55)$_{111}$ (parameter space $N^{\mathscr{X}}$) and (3.69)–(3.70)$_{114}$ (parameter space $\Delta_{\mathscr{X}}$) and the resulting predictive posteriors by (3.50)–(3.52)$_{111}$ (posterior possibility spaces $\imath\check{x}\times\mathscr{X}^{\hat{n}}$, $[\check{m}]\times\mathscr{X}^{\hat{n}}$, $(\mathscr{X}^N)_{\geq\check{m}}$) and (3.65)–(3.68)$_{114}$ (posterior possibility spaces $\imath\check{x}\times\mathscr{X}^{\hat{n}}$, $[\check{m}]\times\mathscr{X}^{\hat{n}}$, $(\mathscr{X}^N)_{\geq\check{m}}$, and $\hat{n}^{\mathscr{X}}$).

(ii) The most interesting conditioning events were the observation of a sample sequence $\check{x}$ and the observation of a count vector $\check{m}$, which corresponds to a permutation class of sample sequences $[\check{m}]$. For updating under both assumptions, if the difference between these observations is irrelevant under the assumptions – i.e., $\check{m} = C_{\mathscr{X}}\check{x}$ –, their likelihood functions are proportional ($L_{\check{x}}\propto L_{\check{m}}$ and $B_{\check{x}}\propto B_{\check{m}}$)

and thus – in concordance with the likelihood principle – the posterior previsions are identical (cf. remarks (i)$_{111}$ and (i)$_{114}$).

We have talked here about the likelihood principle in a context where the original sample and possibility spaces are finite. A finite possibility space $\mathscr{X}^N$ is what we encountered when discussing (finite) exchangeability. But when working with infinitely exchangeable sequences and frequency distributions, the sample space $\mathbb{N}_{>0} \to \mathscr{X}$ becomes countably infinite. However, as recalled just above, the likelihood principle still holds, so we could qualify it as the discrete – instead of just finite – likelihood principle [Walley 1991, §8.6.1–2$_{434–435}$]. Note that the multinomial likelihoods are functions of a continuous argument; this means that the limit arguments we used in the discussion of infinite exchangeability have resulted in a conditional probability defined on an infinite partition!

In classical Bayesian updating, the likelihood principle is further taken to hold for continuous sample spaces as well [see, e.g., Bernardo & Smith 1994, §4.4.1–3$_{181–189}$]. In this case the likelihood function is given more-or-less as a conditional probability density function and Bayes's rule is replaced by a variant for density functions [see, e.g., Walley 1991, §6.10.4$_{331}$]. Walley [1991, §8.6.3–8$_{436–440}$] warns that this continuous likelihood principle can only be relied upon under a potentially very restrictive set of assumptions:

(i) the continuous sample space can be seen as an idealization of a discrete one: measurements have a limited precision;
(ii) the (actual) measurement imprecision is sufficiently small (with relation to the variation in the continuous likelihood function);
(iii) with every continuous sample, a discrete sample can be associated that has a measurement imprecision dependent, positive (discrete) likelihood function;
(iv) the continuous likelihood function is a uniform limit of the suitably normalized discrete likelihood function for decreasing (idealized) measurement imprecision;
(v) the (idealized and thus actual) measurement imprecision is constant over the inferential possibility space (i.e., it does not vary with the likelihood function's argument).

Continuous sample spaces will appear in the next chapter, 'Inference models for exponential families'$_{154}$. So although this warning does not apply here, it will be very relevant there, where we make the – for the scope of applicability – potentially restrictive assumption that the continuous likelihood principle holds.

This whole section, and the updating results of the two previous subsections in particular, provide us with a necessary basis to build concrete inference models on and a context to place them in. These concrete inference models are the subject of the next two sections. The first of these – about predictive inference – builds on the finite exchangeability

assumption; the second – about parametric inference – builds on the infinite exchangeability assumption.

But before diving into those subjects, it is interesting to know the distinction we make between predictive and parametric inference. We follow Geisser [1993]: When the possibility space of the lower prevision used as an inference model consists of observables, we are dealing with predictive inference. When it consists of abstract, mathematical objects that cannot be observed (directly), we are dealing with parametric inference. Quite often, these abstract objects are related to the sample space and observational possibility space by some limiting argument.

It is important to realize that, under this definition, what we called *parametric priors and posteriors* at the beginning of this subsection *can still be predictive inference models*, as long as the parameter space consists of concrete objects. In this section, the first main type of inference models were unconditional and updated sequence distributions defined on possibility spaces $\mathscr{X}^N$ and $\mathscr{X}^{\hat{n}}$ consisting of finite sequences of samples from a sample space $\mathscr{X}$; finite sample sequences are eminently concrete, so these sequence distributions are predictive inference models. The second main type of inference models, prior and posterior count distributions, appeared after the introduction of the finite exchangeability assumption; their sample spaces $N^{\mathscr{X}}$ and $\hat{n}^{\mathscr{X}}$ consist of count vectors, which are concrete as they correspond to (finite sets of) finite sequences of samples, so these count distributions are predictive inference models as well. The third main type, prior and posterior frequency distributions, appeared after the introduction of the infinite exchangeability assumption; their sample space is the set $\Delta_{\mathscr{X}}$ of frequency vectors, both for the prior and the posterior; these frequency vectors are abstract, as they correspond to (infinite sets of) infinite sequences of samples, so frequency distributions are parametric inference models.

3.2 Predictive inference: representation insensitive prediction

This section follows the lines of De Cooman et al. [2007*a*, 2009*b*], but uses our novel definition of exchangeability (cf. §3.1.1$_{95}$), thus avoiding some technical difficulties.

The setting of this section is already quite familiar: we consider a subject who is making a sequence of $N : \mathbb{N}_{>0}$ observations of a certain phenomenon. In the previous section we investigated the consequences of the quite general exchangeability assumption on the form of the sequence distribution and how to update it after having observed a partial sequence of samples.

In this section, we add further assumptions and initially reduce the scope from predictions about all unobserved samples to one unobserved sample. With this, we have a similar aim as the one addressed by the Laplace–Bayes rule of succession [Laplace 1825] and Carnap's [1952] λ-calculus. This section ultimately leads to a predictive inference model for categorical data that is unique up to a parameter that in some sense

determines the speed of learning: the imprecise Dirichlet-multinomial model [Walley & Bernard 1999].

3.2.1 *Immediate prediction: previsions, families & systems*

Conceptually, we start from some nonexplicit exchangeable set of desirable gambles $\mathscr{R}^N_{\mathscr{X}} :\subset \mathscr{L}_{\mathscr{X}^N}$ (cf. §3.1.1$_{95}$), where $\mathscr{X}$ is a finite set of categories (why we add a sub- and superscript will become clear later on). With this set of desirable gambles, there corresponds an (equally unexplicated) joint exchangeable sequence distribution $\underline{P}^N_{\mathscr{X}}$ on $\mathscr{L}_{\mathscr{X}^N}$, which we know is fully characterized through the representation theorem (3.19)$_{102}$ by its coherent prior count distribution $\underline{Q}^N_{\mathscr{X}}$, defined on $\mathscr{L}_{N^{\mathscr{X}}}$.

After the observation of the first $\check{n} : 1..N-1$ samples $\check{x} : \mathscr{X}^{\check{n}}$ with count vector $\check{m} : \check{n}^{\mathscr{X}}$, there remain $\hat{n} := N - \check{n}$ unobserved samples. Our uncertainty about what they turn out to be can be modeled with the (also nonexplicit) updated joint exchangeable sequence distribution $\underline{P}^{\hat{n}}_{\mathscr{X}}(\bullet|\check{x})$ on $\mathscr{X}^{\hat{n}}$ (cf. (3.50)$_{111}$) and its corresponding coherent posterior count distribution $\underline{Q}^{\hat{n}}_{\mathscr{X}}(\bullet|\check{m})$ on $\mathscr{L}_{\hat{n}^{\mathscr{X}}}$ (cf. (3.54)$_{111}$), where $\check{m} := C_{\mathscr{X}}\check{x}$.

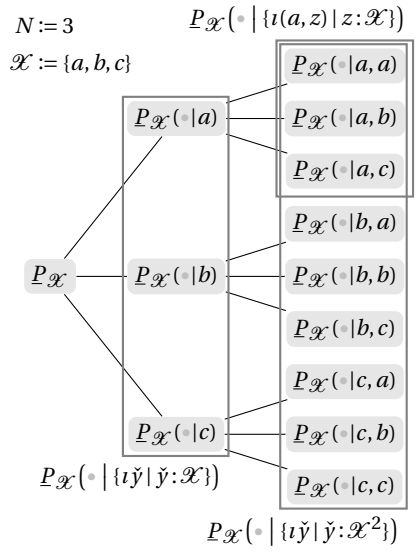
These unexplicated distributions provide models for our uncertainty about *all* the unobserved samples. In this section, we are initially going to restrict our scope to modeling our uncertainty about (any) *one* future unobserved sample; the next sample, for example. The $\mathscr{X}$-marginals of $\underline{P}^N_{\mathscr{X}}$ and $\underline{P}^{\hat{n}}_{\mathscr{X}}(\bullet|\check{x})$ provide the appropriate coherent lower previsions; we respectively denote them by $\underline{P}_{\mathscr{X}}$ and $\underline{P}_{\mathscr{X}}(\bullet|\check{x})$ (so we leave out a superscript 1); they are both defined on $\mathscr{L}_{\mathscr{X}}$. We call them predictive sequence previsions; they are the basic unit in the predictive inference framework we are building. The set of all the $\mathscr{X}$-marginals

$$\sigma_{\mathscr{X}} := \imath\underline{P}_{\mathscr{X}} \cup \bigcup_{\check{n}:1..N-1} \left\{ \underline{P}_{\mathscr{X}}(\bullet|\check{y}) \,\middle|\, \check{y} : \mathscr{X}^{\check{n}} \right\} \tag{3.74}$$

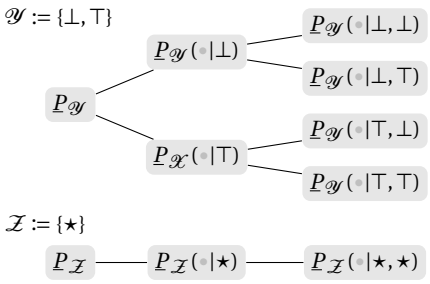
$$:= \imath\underline{P}_{\mathscr{X}} \cup \bigcup_{\check{n}:1..N-1} \underline{P}_{\mathscr{X}}\left(\bullet \,\middle|\, \{\imath\check{y} \,|\, \check{y} : \mathscr{X}^{\check{n}}\}\right) \tag{3.75}$$

is called an $\mathscr{X}$-family of predictive sequence previsions or a predictive $\mathscr{X}$-family for short. The second definition is a reformulation in terms of conditional lower previsions. An example family is given on the side in an illustration using a probability tree [see, e.g., Shafer 1996].

The set of all $\mathscr{X}$-families is denoted by $\Sigma_{\mathscr{X}}$; one $\mathscr{X}$-family differs from another when corresponding member predictive previsions (i.e., having the same conditioning event) have a different value in the same gamble. So predictive families can be partially ordered by pointwise comparison, lower prevision by lower prevision.



We started out with a certain set of categories $\mathscr{X}$. Often, however, there is no unique way of categorizing a set of observations; compare the following two sets of categories for animals, for example: {insect,bird, fish,mammal} versus {flying,swimming,walking}. With every categorization conceivable, we can associate a family of predictive previsions. Let $\mathscr{S}$ denote the collection of all finite sets. A grouping of families for all possible categorizations



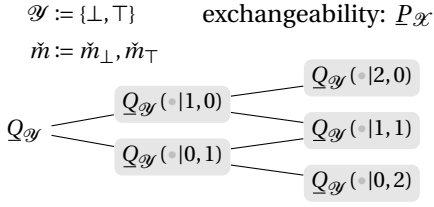
$$\sigma := \{\sigma_{\mathscr{X}} \mid \mathscr{X} : \mathscr{S}\} \tag{3.76}$$

is called a system of predictive previsions or predictive system for short. The example family given earlier can be complemented by other example families to give an abstract impression of what a predictive system looks like.

The set of all predictive systems (for the fixed sample sequence length N) is denoted by Σ. Two predictive systems can be compared family by family and therefore also Σ can be partially ordered.

Now that we know what predictive previsions, families, and systems are, let us see what is the effect of our assumption that – for each predictive family – the predictive sequence previsions are marginals of some (unexplicated) *exchangeable* joint sequence distribution (either $\underline{P}^N_{\mathscr{X}}$ or $\underline{P}^{\hat{n}}_{\mathscr{X}}(\bullet|\check{x})$). To this end, consider some $\mathscr{X}$-family. The respective predictive count previsions (i.e., the induced coherent count distributions) $\underline{Q}_{\mathscr{X}}$ and $\underline{Q}_{\mathscr{X}}(\bullet|\check{m})$, which are both defined on $\mathscr{L}_{1^{\mathscr{X}}}$, are then actually *equivalent* (as uncertainty models) to the corresponding predictive sequence previsions $\underline{P}_{\mathscr{X}}$ and $\underline{P}_{\mathscr{X}}(\bullet|\check{x})$, due to the isomorphism between $\mathscr{X}$ and $1^{\mathscr{X}}$. We therefore consider these predictive count previsions to also be members of an exchangeable predictive family or system if their corresponding predictive sequence previsions are. To boot, we let them be implicitly defined on $\mathscr{L}_{\mathscr{X}}$ as well.

The fact that we can interchange predictive previsions with predictive count previsions reminds us of the following consequence of exchangeability: $\underline{P}_{\mathscr{X}}(\bullet|\check{x})$ only depends on $\check{x}$ through the associated count vector $\check{m}$. This is only the most visible effect of the exchangeability assumption: even though this assumption plays no role for these predictive (count) distributions individually, as the sequence length of interest has been reduced to one, there are relationships between them.



The importance of the fact that 'there are relationships between them' is better appreciated if we turn things around. Up until now, we took a top-down approach: starting from some unexplicated prior set of desirable gambles, we generated a predictive family by formulating

things in terms of previsions, conditioning and marginalizing. However, at times, such a prior is a rather big rabbit to conjure up out of our hat. In a bottom-up approach, we can just conjure up – or, more mundanely put: propose – simple building blocks: the predictive previsions themselves. But now the onus is on us to make sure that these predictive previsions are marginals of some exchangeable joint sequence distribution (or one of the exchangeable updated joint sequence distributions).

Ensuring that the relationships between the predictive previsions are such that they can be seen as marginals of some exchangeable joint prevision – that need not be unique – is a problem for which I am not aware if there is a general solution (yet). So this is something we need to check for every predictive family we propose to use.

Two partial results can ease the pain of our disappointment:

(i) To guarantee the coherence of all these predictive previsions with the unexplicated joint $\underline{P}^N_{\mathscr{X}}$ (cf. §1.3.4$_{59}$), it is enough to make sure that each predictive prevision in the family is coherent (1.28)$_{41}$, or, in other words, to make sure that each conditional prevision in the family is separately coherent (cf. §1.3.4$_{59}$ again). This is a consequence of the marginal extension theorem (1.84)$_{62}$ [also see Miranda & De Cooman 2007, Thm 4].

A predictive family is called coherent if every member predictive prevision is, and a predictive system is called coherent when each member family is. We denote the set of all coherent $\mathscr{X}$-families by $(\Sigma_{\mathscr{X}})_{\text{coh}}$ and the set of all coherent systems by Σ_{coh}.

(ii) We can obtain a necessary (but not sufficient) condition for exchangeability that will be useful later on. Consider two 'successive' predictive previsions $\underline{P}_{\mathscr{X}}(\bullet|\check{x})$ and $\underline{P}_{\mathscr{X}}\big(\bullet \,\big|\, \{\iota(\check{x},z) \,|\, z:\mathscr{X}\}\big)$, where $\check{x}$ is either empty – in which case $\underline{P}_{\mathscr{X}}(\bullet|\check{x})$ reduces to $\underline{P}_{\mathscr{X}}$ –, or an element of $\{y:\mathscr{X}^* \,|\, \nu y < N-1\}$. To possibly be (updated) marginals of an exchangeable joint $\underline{P}^N_{\mathscr{X}}$, they must respectively be an $\mathscr{X}$-marginal and a conditional prevision of its updated exchangeable $\mathscr{X}^2$-marginal $\underline{P}^2_{\mathscr{X}}(\bullet|\check{x})$. This $\mathscr{X}^2$-marginal dominates the marginal extension $\underline{P}_{\mathscr{X}}\big(\underline{P}_{\mathscr{X}}(\bullet|\check{x},\bullet) \,\big|\, \check{x}\big)$ by definition (1.84)$_{62}$, as that is the least committal (coherent) extension and does not take exchangeability into account. Now let f be some gamble on $\mathscr{X}$; by using cylindrical extension (cf. §1.3.1$_{52}$) and by explicitly showing which random variable takes up which argument position, we can write

$$\underline{P}_{\mathscr{X}}(f|\check{x}) = \underline{P}_{\mathscr{X}}\big(f(X_{\check{n}+1}) \,\big|\, \check{x}\big) = \underline{P}^2_{\mathscr{X}}\big(\tilde{f}(X_{\check{n}+1}, X_{\check{n}+2}) \,\big|\, \check{x}\big);$$

next, we use the permutation invariance of $\underline{P}^2_{\mathscr{X}}(\bullet|\check{x})$ that is implied by its exchangeability (cf. (3.50)$_{111}$, (3.19)$_{102}$,and (3.7)$_{99}$):

$$\begin{aligned} &= \underline{P}^2_{\mathscr{X}}\big(\tilde{f}(X_{\check{n}+2}, X_{\check{n}+1}) \,\big|\, \check{x}\big) \\ &= \underline{P}^2_{\mathscr{X}}\big(f(X_{\check{n}+2}) \,\big|\, \check{x}\big); \end{aligned}$$

then, we use the dominance of the $\mathscr{X}^2$-marginal over the marginal extension and finish by eliminating the cylindrical extension:

$$\begin{aligned} &\geq \underline{P}_{\mathscr{X}}\Big(\underline{P}_{\mathscr{X}}\big(f(X_{\check{n}+2}) \,\big|\, \check{x}, X_{\check{n}+1}\big) \,\Big|\, \check{x}\Big) \\ &= \underline{P}_{\mathscr{X}}\big(\underline{P}_{\mathscr{X}}(f|\check{x}, \bullet) \,\big|\, \check{x}\big), \end{aligned} \tag{3.77}$$

A predictive family is called exchangeable if its member predictive previsions are marginals of some exchangeable (updated) joint and a predictive system is exchangeable when each member family is. We denote the set of all exchangeable $\mathscr{X}$-families by $(\Sigma_{\mathscr{X}})_{\text{xch}}$ and the set of all exchangeable systems by Σ_{xch}. (Recall that our exchangeability definition $(3.4)_{97}$ presupposes and as such includes coherence.)

In this subsection, we have introduced the concept of systems of predictive previsions, but we still seem far removed from proposals for concrete predictive systems that can be used in practice. The main problem is that we have too much choice: requiring exchangeability (and thus coherence) provides rather few restrictions on the values we can assign to predictive previsions and, more importantly, almost no guidance on how to assign them. The art of creating inference models is a difficult one if there is too much freedom. We need more restrictions than only exchangeability; we need guiding principles to bring order to the land of predictive systems! In the next subsection, we provide one such principle.

3.2.2 *Representation insensitivity*

Halpern & Koller [2004] provide a good list of references on representation (in)dependence.

I do not think there are universally valid principles, only principles that are acceptable within a certain context. So if we want to follow some principle in designing predictive systems, we need to sketch the context within which this principle is reasonable. We have already assumed that the order of the samples is irrelevant (exchangeability). Now we additionally assume that

(i) we start from a state of complete prior ignorance, i.e., we dare not even hazard a guess as to which categories will be observed;

(ii) the sample data really are categorical in the sense that they do not belong to a set with any – e.g., ordinal – structure that cannot be preserved under arbitrary recategorization of the data.

These assumptions, which are reasonable or reasonable approximations in many practical situations (e.g., for an analyst without any domain knowledge about the dataset he has been given), allow us to propose three invariance principles that the predictive systems should satisfy.

The first principle is pooling invariance: because we do not know how the samples will be categorized, our uncertainty model should not intrinsically change when we group together (pool) some categories to form a new category. Mathematically, we can formalize this idea as

follows: Consider a set of categories $\mathscr{X}$ and a partition $\mathscr{Z}$ of this set $\mathscr{X}$; this partition $\mathscr{Z}$ can be seen as an alternative categorization formed by pooling some categories of $\mathscr{X}$ together. Let $\varphi : \mathscr{X} \to \mathscr{Z}$ be the surjective map corresponding to this partition (cf. §1.3.1$_{52}$). Now consider some gamble f on $\mathscr{Z}$, then we should have that $\underline{P}_{\mathscr{X}}(f \circ \varphi | \check{x}) = \underline{P}_{\mathscr{Z}}(f | \varphi \check{x})$: for gambles that do not differentiate between original categories in the same pool, it should not matter whether we consider predictive inferences for the set of original categories $\mathscr{X}$, or for the set of pooled categories $\mathscr{Z}$.

Our pooling principle corresponds to (and is perhaps a more apt name for) Walley's [1996] representation invariance principle.

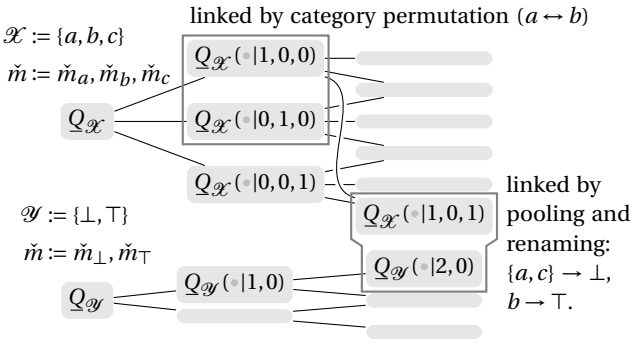
The second principle is renaming invariance: as long as no confusion can arise, it should not matter for a subject's predictive inferences what names he gives to the different categories. This may seem too trivial to even mention, and as far as we know, it is always implicitly taken for granted in predictive inference. But it will be well to devote some attention to it here, in order to distinguish it from the category permutation invariance to be discussed shortly, with which it is easily confused if we do not pay proper attention. Mathematically, we can formalize it as follows: Consider two isomorphic but distinct sets of categories $\mathscr{X}$ and $\mathscr{Y}$ between which we can define a renaming bijection $\psi : \mathscr{X} \leftrightarrow \mathscr{Y}$ that either leaves an element's name intact or replaces it by a name not present in the original set. Now consider some gamble f on $\mathscr{Y}$, then we should have that $\underline{P}_{\mathscr{X}}(f \circ \psi | \check{x}) = \underline{P}_{\mathscr{Y}}(f | \psi \check{x})$.

The third principle is permutation invariance: in a state of prior ignorance, a subject has no reason to distinguish between the different elements of any set of categories $\mathscr{X}$ he might use. Mathematically, we can formalize it as follows: Consider any permutation $\pi : \Pi_{\mathscr{X}}$ of the categories. With any gamble f on $\mathscr{X}$, there corresponds a pointwise permuted gamble $f \circ \pi$ and with the observed sequence $\check{x}$ there corresponds a permuted sequence $\pi\check{x}$. If a subject has no reason to make a distinction between a category and its permutation, then we should have that $\underline{P}_{\mathscr{X}}(f \circ \pi | \check{x}) = \underline{P}_{\mathscr{X}}(f | \pi \check{x})$.

We encountered (weak) permutation invariance before, in §2.2.6$_{83}$; it should not be confused with strong permutation invariance (e.g., exchangeability §3.1.1$_{95}$).

As both ψ and π are bijections, renaming invariance and permutation invariance are very similar. But ψ cannot even partly act like a permutation: it does not allow confusion between original and renamed categories. However, it is possible that two subsequent renamings result in a permutation, so it seems rather difficult to mathematically separate the two concepts. Luckily, this has never been our intention.

It is our intention to consider all three invariance principles simultaneously. We call this combination the representation insensitivity principle. Its mathematical formalization is a straightforward generalization: Let $\mathscr{X}$ and $\mathscr{Y}$ be any two finite sets of categories that can be related to each-other with a surjective relabeling map $\rho : \mathscr{X} \to \mathscr{Y}$ and let f be a gamble on $\mathscr{Y}$, then the representation insensitivity of a predictive system implies the following relation between two of its predictive sequence previsions: $\underline{P}_{\mathscr{X}}(f \circ \rho | \check{x}) = \underline{P}_{\mathscr{X}}(f | \rho \check{x})$.



Representation insensitivity and the invariances it subsumes can of course also be expressed in terms of predictive count previsions. Actually, we will use predictive count previsions to formulate our central definition of representation insensitivity, as they already partially express exchangeability.

To arrive at this definition, we first of all need to translate sequence relabelings to count vector relabelings. This is done as follows: We take the same relabeling map ρ from $\mathscr{X}$ to $\mathscr{Y}$ as before, then the count vector $\check{m}$ on $\check{n}^{\mathscr{X}}$ is transformed into a count vector $C_\rho \check{m}$ on $\check{n}^{\mathscr{Y}}$ defined for every z in $\mathscr{Y}$ by $(C_\rho \check{m})_z := \sum_{x:\mathscr{X}\,\wedge\,\rho x = z} \check{m}_x$, i.e., we sum up the counts for the categories that have the same label after relabeling. As a second step towards our central definition, consider that any gamble f on any set of categories $\mathscr{X}$ can be seen as a relabeling map with its own range $\{f\}$ as the new set of labels.

We can then combine both steps and use representation insensitivity to rewrite the predictive count prevision of the gamble f in a so-called standard representation insensitive form, i.e., as part of the exchangeable $\{f\}$-family:

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) = \underline{Q}_{\{f\}}(\mathrm{id}_{\{f\}}|C_f\check{m}). \tag{3.78}$$

This standard representation insensitive form shows that $\underline{Q}_{\mathscr{X}}(f|\check{m})$, *and thus also* $\underline{P}_{\mathscr{X}}(f|\check{x})$, *only depend on the values that* f *may assume, and on the number of times each value has been observed.*

Our central definition of representation insensitivity of a predictive system then expresses that two previsions that have the same standard representation insensitive form must have the same value. This is formalized by the predicate $\mathrm{rip}:\Sigma \to \mathbb{B}$ with

$$\begin{aligned}
&\mathrm{rip}\,\sigma \Leftrightarrow \forall \mathscr{X}, \mathscr{Y} : \mathscr{S}^2;\\
&\qquad \forall f, g : \mathscr{L}_{\mathscr{X}} \times \mathscr{L}_{\mathscr{Y}} \wedge \{f\} = \{g\};\\
&\qquad \forall \underline{Q}_{\mathscr{X}}(\bullet|\check{m}), \underline{Q}_{\mathscr{Y}}(\bullet|\check{m}') : \sigma_{\mathscr{X}} \times \sigma_{\mathscr{Y}} \wedge C_f\check{m} = C_g\check{m}';\\
&\qquad\quad \underline{Q}_{\mathscr{X}}(f|\check{m}) = \underline{Q}_{\mathscr{Y}}(g|\check{m}').
\end{aligned} \tag{3.79}$$

We can denote the set of all (coherent or exchangeable) representation insensitive predictive systems by Σ_{rip} ($\Sigma_{\mathrm{coh}\wedge\mathrm{rip}}$ or $\Sigma_{\mathrm{xch}\wedge\mathrm{rip}}$). In contrast to what happens for exchangeability (cf. $(3.6)_{99}$), systems that dominate a representation insensitive system need not be representation insensitive.

3.2.3 *Properties of representation insensitive predictive systems*

Definition (3.79) provides us with a compact characterization of representation insensitivity that underlines its basic meaning. It does not provide us with immediate insight into what its consequences are for coherent predictive systems. To get this insight, we investigate here some consequences for the predictive lower and upper probabilities implied by the member predictive previsions.

Consider some *proper* subset A of $\mathscr{X}$, then (3.78) allows us to write

$$\underline{Q}_{\mathscr{X}}(A|\check{m}) = \underline{Q}_{\{0,1\}}\big(\mathrm{id}_{\{0,1\}} \,\big|\, \check{n} - \textstyle\sum \check{m}_A, \sum \check{m}_A\big), \tag{3.80}$$

$$\begin{aligned}\overline{Q}_{\mathscr{X}}(A|\check{m}) &= 1 - \underline{Q}_{\mathscr{X}}(\mathscr{X} \setminus A|\check{m}) \\ &= 1 - \underline{Q}_{\{0,1\}}\big(\mathrm{id}_{\{0,1\}} \,\big|\, \textstyle\sum \check{m}_A, \check{n} - \sum \check{m}_A\big).\end{aligned} \tag{3.81}$$

We see from these expressions that *the predictive lower and upper probabilities only depend on the total number of observed samples $\check{n}$ and the number $\sum \check{m}_A$ of them that belong to A*. Using the so-called lower probability function $q : \{k, \ell : (\mathbb{N}_{<N})^2 \mid \ell \leq k\} \to [0,1]$ of the predictive system, we can compactly write these probabilities; it is defined by

$$q(k,\ell) = \underline{Q}_{\{0,1\}}(\mathrm{id}_{\{0,1\}} \,|\, k-\ell, \ell). \tag{3.82}$$

This restriction on the probability values is similar to (but stronger than) Johnson's sufficientness postulate [Zabell 1982].

It gives the lower probability of seeing an event that has been observed ℓ times in k trials. Whenever it is needed to avoid ambiguity, we add a superscript to q to indicate which predictive system it refers to.

The fact that the range of q is $[0,1]$-bounded follows from the preassumed coherence of the predictive system; i.e., from accepting sure gains $(1.29)_{42}$ (lower bound) and from normedness $(1.34)_{42}$ and $(1.33)_{42}$ (upper bound).

Another consequence of coherence that can be translated to the lower probability function is superadditivity $(1.31)_{42}$. For this, consider the set $\{a,b,c\}$ and assume that we observed a sequence with count vector $\check{m} = (\check{m}_a, \check{m}_b, \check{m}_c) = (\ell', \ell'', \check{n} - \ell' - \ell'')$, where ℓ' and ℓ'' in $\mathbb{N}_{\leq \check{n}}$ are such that $\ell' + \ell'' \leq \check{n}$, then

$$\begin{aligned}q(\check{n}, \ell' + \ell'') &= \underline{Q}_{\{a,b,c\}}\big(\{a,b\} \,\big|\, \check{m}\big) \\ &\geq \underline{Q}_{\{a,b,c\}}(a|\check{m}) + \underline{Q}_{\{a,b,c\}}(b|\check{m}) = q(\check{n},\ell') + q(\check{n},\ell'').\end{aligned}$$

Remember that due to representation insensitivity, it is enough to prove the property using any one category set; the one used here was chosen for its simplicity. So for all k in $\mathbb{N}_{<N}$ and all ℓ', ℓ'' in $\mathbb{N}_{\leq k}$ such that $\ell' + \ell'' \leq k$ we have

$$q(k, \ell' + \ell'') \geq q(k,\ell') + q(k,\ell''). \tag{3.83}$$

Some immediate consequences of the second-argument superaddi-

tivity are: (now let $\ell : \mathbb{N}_{\leq k}$ and $\ell' : \mathbb{N}_{<k}$)

$$q(k,0) = 0, \tag{3.84}$$

$$q(k,\ell) \geq \ell \cdot q(k,1) \quad \text{and} \quad q(k,1) \leq \tfrac{1}{k}, \tag{3.85}$$

$$q(k,\ell'+1) \geq q(k,\ell'). \tag{3.86}$$

Two interesting and intuitively appealing conclusions about the predictive lower and upper probabilities that are valid in any representation insensitive coherent predictive system follow from these:

(i) Equation (3.84) tells us that the lower probability of observing an event that has not been observed before is zero; by conjugacy, it similarly holds that the upper probability of observing an event that has always been observed before is one.

(ii) Equation (3.86) tells us that, for a fixed total number of observations, both the lower and the upper probability of some event do not decrease if the number of times that event has been observed increases.

The first of these conclusions has a quite far-reaching consequence: for any representation invariant coherent predictive system, the initial predictive prevision of any of its $\mathscr{X}$-families must be vacuous; i.e., $\underline{P}_{\mathscr{X}} = \min$ (and $\underline{Q}_{\mathscr{X}} = \min$). To see this, consider any nonconstant gamble f on $\mathscr{X}$, then

$$\begin{aligned}
0 &\leq \underline{Q}_{\mathscr{X}}\big(f - \min\{f\}\big) \\
&= \underline{Q}_{\{f\}}\big(\mathrm{id}_{\{f\}} - \min\{f\}\big) \\
&\leq \big(\max\{f\} - \min\{f\}\big) \cdot \underline{Q}_{\{f\}}\{f\}_{>\min\{f\}} + 0 \cdot \overline{Q}_{\{f\}}\{f\}_{=\min\{f\}} \\
&= \big(\max\{f\} - \min\{f\}\big) \cdot q(0,0) = 0,
\end{aligned}$$

where the first inequality follows from accepting sure gains $(1.29)_{42}$ and the second from mixed subadditivity $(1.36)_{42}$. So then constant additivity $(1.35)_{42}$ allows us to conclude $\underline{Q}_{\mathscr{X}} f = \min\{f\}$; for constant gambles this result immediately follows from normedness $(1.34)_{42}$.

Up until now, we have only assumed the predictive system under scrutiny to be coherent. If we additionally assume it to be exchangeable, we have the inequality $(3.77)_{122}$ we found earlier (and its reformulation in terms of predictive count previsions) at our disposal as a restriction. So consider the set $\{a,b\}$ and assume that we have observed a sequence of length $\check{n} : \mathbb{N}_{<N-1}$ with count vector $\check{m} = (\check{m}_a, \check{m}_b) = (\ell, \check{n} - \ell)$, where $\ell : \mathbb{N}_{<\check{n}}$, then we infer from $(3.77)_{122}$ that

$$\begin{aligned}
q(\check{n},\ell) &= \underline{Q}_{\{a,b\}}(a|\check{m}) \\
&\geq \underline{Q}_{\{a,b\}}\Big(\underline{Q}_{\{a,b\}}\big(a \,\big|\, \check{m} + (1,0)\big) \cdot I^a + \underline{Q}_{\{a,b\}}\big(a \,\big|\, \check{m} + (0,1)\big) \cdot I^b \,\Big|\, \check{m}\Big) \\
&= \underline{Q}_{\{a,b\}}\big(q(\check{n}+1,\ell+1) \cdot I^a + q(\check{n}+1,\ell) \cdot I^b \,\big|\, \check{m}\big)
\end{aligned}$$

$$= Q_{\{a,b\}}\big(q(\check{n}+1,\ell) + \big(q(\check{n}+1,\ell+1) - q(\check{n}+1,\ell)\big)\cdot I^a \,\big|\, \check{m}\big)$$

and as we know from $(3.86)_{\frown}$ that $q(\check{n}+1,\ell+1)-q(\check{n}+1,\ell)$ is nonnegative, constant additivity $(1.35)_{42}$ and nonnegative homogeneity $(1.30)_{42}$ allow us to write

$$\begin{aligned}&= q(\check{n}+1,\ell) + \big(q(\check{n}+1,\ell+1) - q(\check{n}+1,\ell)\big)\cdot Q_{\{a,b\}}(a|\check{m})\\ &= q(\check{n}+1,\ell) + \big(q(\check{n}+1,\ell+1) - q(\check{n}+1,\ell)\big)\cdot q(\check{n},\ell).\end{aligned}$$

So for all k in $\mathbb{N}_{<N-1}$ and all ℓ in $\mathbb{N}_{\leq k}$ we have

$$q(k,\ell) \geq q(k+1,\ell) + q(k,\ell)\cdot\big(q(k+1,\ell+1) - q(k+1,\ell)\big). \tag{3.87}$$

There is one interesting immediate consequence of this inequality. We already know that the difference of the second right-hand side term is nonnegative, so we can drop that term and find

$$q(k,\ell) \geq q(k+1,\ell). \tag{3.88}$$

This gives us a third intuitively appealing conclusion, now about the predictive lower and upper probabilities in any exchangeable representation insensitive predictive system: The lower probability for an event for which there are a fixed number of observations does not increase when the total number of observations increases.

All but one property of predictive systems we have encountered in this subsection has been expressed in terms of the lower probability function. These properties are thus geared more towards predictive systems that can be specified entirely in terms of lower and upper probabilities. Such a specification would result in predictive systems consisting of member predictive previsions that are 2-monotone (cf. the third sidenote of §2.2.3$_{72}$). In fact, the concrete predictive systems we are going to propose after the next subsection come from an even more restricted class: those with predictive previsions that are linear-vacuous mixtures (cf. the first sidenote of §2.2$_{70}$).

3.2.4 *The vacuous & Haldane predictive systems*

The fact – discovered in the last subsection – that any representation invariant coherent predictive system must have vacuous initial predictive previsions, when combined with a profound lack of inspiration, results in our first proposal: the vacuous predictive system $\sigma^{\min}$. Its defining property is that *all* member predictive previsions are vacuous. Put otherwise, this means that its lower probability function $q^{\min}$ is identically zero. It is dominated by all other coherent predictive systems.

Does this conservative system par excellence satisfy our requirements of exchangeability and representation insensitivity? The answer is yes on both counts:

(i) Representation insensitivity follows from (let $\underline{Q}_{\mathscr{X}}(\bullet|\check{m})$ be a member of the $\mathscr{X}$-family $\sigma^{\min}_{\mathscr{X}}$ and f some gamble on $\mathscr{X}$)

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) = \min\{f\} = \min\{\mathrm{id}_{\{f\}}\} = \underline{Q}_{\{f\}}\big(\mathrm{id}_{\{f\}} \,\big|\, C_f\check{m}\big).$$

(ii) Exchangeability follows from the fact that, for each $\mathscr{X}$-family, the predictive previsions can be seen as marginals of
   (a) the exchangeable prior joint sequence distribution $\underline{P}^N_{\mathscr{X}}$ that is fully determined by the count distribution $\underline{Q}^N_{\mathscr{X}} = \min$ through the representation theorem (3.19)$_{102}$, and
   (b) the exchangeable updated sequence distributions $\underline{P}^{\hat{n}}_{\mathscr{X}}(\bullet|\check{x})$ obtained via natural extension (3.50)$_{111}$, which are in turn fully determined by their count distributions $\underline{Q}^N_{\mathscr{X}} = \min$ (cf. (3.53)$_{111}$).

In the vacuous predictive system, no learning takes place: starting from a state of prior ignorance, we stay in that state when using this system. Whatever sequence of samples is observed and whatever its length, no new inferences are drawn. Of all the coherent (and thus also of all the exchangeable) predictive systems, it is the least committal one, in the sense that all other coherent predictive systems dominate it.

At the other end of the spectrum are the maximally committal coherent or exchangeable predictive systems, of which we can conjure up one out of our hat. In the realm of coherent previsions, the linear ones are the maximally committal ones (cf. §1.2.8$_{48}$); so what we are going to do is propose a predictive system that consists almost entirely of predictive *linear* previsions.

We know that the initial predictive previsions of every $\mathscr{X}$-family in the proposed system $\sigma^{\max}$ must be vacuous. For all the other predictive previsions, we take a weighted average, where the weight for every category is the frequency of occurrence of that category in the observed sample. So we propose that for any $\underline{Q}_{\mathscr{X}}(\bullet|\check{m})$ in $\sigma^{\max}_{\mathscr{X}}$ for which $\check{n} > 0$ (let f be a gamble on $\mathscr{X}$)

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) = \mathrm{Wa}(f|\check{m}), \tag{3.89}$$

where we have used the weighted average linear prevision. Let $\alpha : (\mathbb{R}^{\mathscr{X}})_{\geq 0}$ such that $\sum\alpha > 0$ be a vector of weights, then this prevision is defined for any gamble f on $\mathscr{X}$ by

$$\mathrm{Wa}(f|\alpha) := \tfrac{1}{\sum\alpha}\cdot\sum f\cdot\alpha = \sum f\cdot\tfrac{\alpha}{\sum\alpha}. \tag{3.90}$$

As can be seen from the last expression, the weights' scale does not matter; i.e., $\mathrm{Wa}(\bullet|\lambda\cdot\alpha) = \mathrm{Wa}(\bullet|\alpha)$ for any positive real λ.

Using a weighted average is a very common idea. We immediately see that the corresponding lower probability function $q^{\max}$ resulting from (3.89) is defined for every k, ℓ in $\mathbb{N}_{<N}$ that are such that $\ell \leq k > 0$ by $q^{\max}(k,\ell) = \frac{\ell}{k}$. This coincides with the inferences from classical frequentist estimation for the (lower and upper) probability of an event

that has been observed ℓ times in k trials. It also coincides with the inferences resulting from a classical Bayesian model with a multinomial likelihood when using Haldane's improper prior [see, e.g., Jeffreys 1983, §III.3.1$_{123}$]. Because of this, we call $\sigma^{\max}$ the Haldane predictive system. No other coherent predictive system dominates it.

But, does this maximally committal, classical system par excellence satisfy our requirements of exchangeability and representation insensitivity? Again, the answer is again yes on both counts:

(i) Representation insensitivity follows from (let $\check{n}:1..N-1$, $\check{m}:\check{n}^{\mathscr{X}}$, and $f:\mathscr{L}_{\mathscr{X}}$)

$$\mathrm{Wa}(f|\check{m}) = \sum f \cdot \tfrac{\check{m}}{\check{n}} = \sum_{r:\{f\}} r \cdot \tfrac{(C_f\check{m})_r}{\check{n}} = \mathrm{Wa}\big(\mathrm{id}_{\{f\}} \,\big|\, C_f\check{m}\big).$$

The Haldane system is also the unique representation insensitive system with only predictive *linear* previsions for $\check{n}>0$. To see this, return to (3.83)$_{125}$ and its consequence (3.85)$_{126}$; for $k>0$ their inequalities become equalities for systems with only predictive linear previsions for $\check{n}>0$, which makes the lower probability function of the Haldane system the only possible one.

(ii) Showing that the Haldane system is exchangeable requires a bit more work, but it will be worth it in terms of interesting side discoveries. To be exchangeable, the predictive previsions should be marginals of an exchangeable joint sequence distribution or an exchangeable updated joint sequence distribution. Because, for every family, all but one of the predictive previsions is linear, calculating these joints is a straightforward (but not simple) iterated application of the marginal extension theorem (1.84)$_{62}$; i.e., a concatenation of the marginals.

To get some feeling for what this concatenation entails, we are first going to look at the situation where $\check{n}:=N-2$; again $\check{x}:\mathscr{X}^{\check{n}}$ and, as always, $\check{m}:=C_{\mathscr{X}}\check{x}$. We let f be a gamble on $\mathscr{X}^2$, then

$$\begin{aligned}\underline{P}^2_{\mathscr{X}}(f|\check{x}) &= \mathrm{Wa}\big(\mathrm{Wa}(f|\check{m}+C_{\mathscr{X}}{}_{\bullet}) \,\big|\, \check{m}\big) && (3.91)\\ &= \sum_{\hat{y}_{N-1}:\mathscr{X}}\Big(\sum_{\hat{y}_N:\mathscr{X}} f(\hat{y}_{N-1},\hat{y}_N)\cdot\frac{(\check{m}+C_{\mathscr{X}}\hat{y}_{N-1})_{\hat{y}_N}}{N-1}\Big)\cdot\frac{\check{m}_{\hat{y}_{N-1}}}{N-2} \\ &= \sum_{\hat{y}:\mathscr{X}^2} f\hat{y}\cdot\frac{\check{m}_{\hat{y}_1}\cdot(\check{m}_{\hat{y}_2}+\delta_{\hat{y}_1\hat{y}_2})}{(N-2)\cdot(N-1)}. && (3.92)\end{aligned}$$

Introducing the Kronecker delta: $\delta_{ab} = I^a b = I^b a$.

Note the invariance of the second factor under permutation of the sequence order. This formula can be immediately generalized to the case where $\check{n}$ is any number in $1..N-1$: (now let f be a gamble on $\mathscr{X}^{\hat{n}}$)

$$\underline{P}^{\hat{n}}_{\mathscr{X}}(f|\check{x}) = \sum_{\hat{y}:\mathscr{X}^{\hat{n}}} f\hat{y}\cdot\prod_{i:1..\hat{n}}\frac{\check{m}_{\hat{y}_i}+\sum_{j:1..i-1}\delta_{\hat{y}_j\hat{y}_i}}{\check{n}-1+i} \qquad (3.93)$$

Recall from §0.3.1$_{24}$ that an interval with an upper bound that is strictly lower than its lower bounds corresponds to the empty set $\emptyset$.

To prove that this formula is correct, we add an inductive step to

the two-step concatenation above by again using the marginal extension theorem $(1.84)_{62}$:

$$\begin{aligned}
\mathrm{Wa}\big(\underline{P}^{\hat{n}-1}_{\mathscr{X}}(f|\check{x},\bullet)\,\big|\,\check{m}\big) &= \textstyle\sum_{\hat{y}_0:\mathscr{X}} \underline{P}^{\hat{n}-1}_{\mathscr{X}}(f|\check{x},\hat{y}_0)\cdot\frac{\check{m}_{\hat{y}_0}}{\check{n}}\\
&= \textstyle\sum_{\hat{y}_0:\mathscr{X}}\Big(\sum_{\hat{y}:\mathscr{X}^{\hat{n}-1}} f(\hat{y}_0,\hat{y})\\
&\qquad\qquad \cdot\prod_{i:1..\hat{n}-1}\frac{\check{m}_{\hat{y}_i}+\delta_{\hat{y}_0\hat{y}_i}+\sum_{j:1..i-1}\delta_{\hat{y}_j\hat{y}_i}}{(\check{n}+1)-1+i}\Big)\cdot\frac{\check{m}_{\hat{y}_0}}{\check{n}}\\
&= \textstyle\sum_{\hat{y}_0:\mathscr{X}}\sum_{\hat{y}:\mathscr{X}^{\hat{n}-1}} f(\hat{y}_0,\hat{y})\cdot\prod_{i:0..\hat{n}-1}\frac{\check{m}_{\hat{y}_i}+\sum_{j:0..i-1}\delta_{\hat{y}_j\hat{y}_i}}{\check{n}+i}\\
&= \underline{P}^{\hat{n}}_{\mathscr{X}}(f|\check{x}).
\end{aligned}$$

Now let us rewrite $(3.93)_{\frown}$:

$$\begin{aligned}
\underline{P}^{\hat{n}}_{\mathscr{X}}(f|\check{x}) &= \textstyle\sum_{\hat{y}:\mathscr{X}^{\hat{n}}} f\,\hat{y}\cdot\prod_{i:1..\hat{n}}\frac{\check{m}_{\hat{y}_i}+\sum_{j:1..i-1}\delta_{\hat{y}_j\hat{y}_i}}{\check{n}-1+i}\\
&= \textstyle\sum_{\hat{y}:\mathscr{X}^{\hat{n}}} f\,\hat{y}\cdot\frac{\prod_{z:\mathscr{X}}\prod_{i:1..\hat{n}\wedge\hat{y}_i=z}\big(\check{m}_z+|\{j:1..i-1\mid\hat{y}_j=z\}|\big)}{\prod_{i:1..\hat{n}}(\check{n}-1+i)};
\end{aligned}$$

for sequences $\hat{y}$ that contain observations z for which $\check{m}_z=0$, the numerators are zero, so we can restrict attention to $\mathscr{X}_{\check{m}}$; splitting the sequence sum into a count vector sum and an atom-restricted sequence sum, we get

$$\begin{aligned}
&= \textstyle\sum_{\hat{m}:\hat{n}^{\mathscr{X}_{\check{m}}}}\big(\sum_{\hat{y}:[\hat{m}]} f\,\hat{y}\big)\cdot\frac{\prod_{z:\mathscr{X}_{\check{m}}}\prod_{i:0..\hat{m}_z-1}(\check{m}_z+i)}{\prod_{i:1..\hat{n}}(\check{n}-1+i)}\\
&= \textstyle\sum_{\hat{m}:\hat{n}^{\mathscr{X}_{\check{m}}}}\big(\sum_{\hat{y}:[\hat{m}]} f\,\hat{y}\big)\cdot\frac{\prod_{z:\mathscr{X}_{\check{m}}}(\check{m}_z+\hat{m}_z-1)!/(\check{m}_z-1)!}{(N-1)!/(\check{n}-1)!};
\end{aligned}$$

using binomial coefficients and $(3.12)_{100}$, this becomes

$$\begin{aligned}
&= \tfrac{1}{\binom{N-1}{\hat{n}}}\cdot\textstyle\sum_{\hat{m}:\hat{n}^{\mathscr{X}_{\check{m}}}}\big(\frac{1}{|[\hat{m}]|}\cdot\sum_{\hat{y}:[\hat{m}]} f\,\hat{y}\big)\cdot\prod_{z:\mathscr{X}_{\check{m}}}\tbinom{\check{m}_z+\hat{m}_z-1}{\hat{m}_z}\\
&= \mathrm{Dm}\big(\mathrm{Mh}(f|\bullet)\,\big|\,\hat{n},\check{m}\big), \qquad (3.94)
\end{aligned}$$

where, in the last step, we spotted the multivariate hypergeometric prevision (cf. $(3.18)_{101}$) and introduced the prevision corresponding to the multivariate negative hypergeometric distribution [Johnson et al. 1997, §39.4.1$_{179}$], which we will encounter again later when we define it under the name of Dirichlet-multinomial prevision $(3.106)_{135}$. The exchangeability of $\underline{P}^{\hat{n}}_{\mathscr{X}}(f|\check{x})$ is evident from the expression obtained (cf. $(3.19)_{102}$).

We only need to check the case where no observations have yet been made; i.e., we need to check if the prior is exchangeable. Using the marginal extension theorem $(1.84)_{62}$ again, we find (now

let f be a gamble on $\mathscr{X}^N$)

$$\begin{aligned}\underline{P}^N_{\mathscr{X}} f &= \underline{P}_{\mathscr{X}}\big(\underline{P}^{N-1}_{\mathscr{X}}(f|\bullet)\big)\\ &= \min_{z:\mathscr{X}} \underline{P}^{N-1}_{\mathscr{X}}\big(f(z,\bullet)\,\big|\,z\big)\\ &= \min_{z:\mathscr{X}} \mathrm{Dm}\Big(\mathrm{Mh}\big(f(z,\bullet)\,\big|\,\bullet\big)\,\Big|\,N-1, C_{\mathscr{X}} z\Big);\end{aligned}$$

now, as $\mathscr{X}_{C_{\mathscr{X}} z} = \imath z$, this can be rewritten as

$$\begin{aligned}&= \min_{y:\{(\imath z)^N | z:\mathscr{X}\}} f y\\ &= \min_{m:N^{\mathscr{X}} \wedge |[m]|=1} \mathrm{Mh}(f|m),\end{aligned}$$

so the prior $\underline{P}^N_{\mathscr{X}}$ is clearly exchangeable. In the third step, we also see that the updated prevision $\underline{P}^{N-1}_{\mathscr{X}}(\bullet|\check{x})$ after one observation $\check{x}:\mathscr{X}$ is extremely committal: it states that we can be practically sure that only the observed category $\check{x}$ will ever be observed again!

Most updated previsions from this joint, obtained using natural or regular extension, would be vacuous. However, in this context, they are not: we started out with them and used them to obtain this joint, with which they are coherent by construction (cf. $(\mathrm{i})_{121}$).

So, in the Haldane predictive system, learning does take place, but it seems to be tied very strongly to the observations: the possibility of getting a sample of a category that has not been observed before is not taken into account.

For people who love to go from one extreme to the next, this subsection has been a real treat. Everybody else will also be happy, happy to know that the two predictive systems we have encountered here – the vacuous $\sigma^{\min}$ and Haldane $\sigma^{\max}$ –, academic as they may be, form an excellent foundation to build really practical predictive systems. This is the subject of the next subsection.

3.2.5 *Mixing predictive systems & the imprecise Dirichlet-multinomial model*

One way to leave the land of extremes and return to our beloved land of compromise is to combine the extremes. We can create a whole class of predictive systems by taking convex mixtures of the vacuous and the Haldane predictive systems. With this, we will be able to obtain predictive systems that are exchangeable, representation insensitive, that learn, and, while doing so, still keep the possibility of observing as yet unobserved categories open.

These convex mixtures are defined as follows: Consider a so-called mixing sequence ε in $[0,1]^{N-1}$, then the mixing predictive system σ^ε is defined by taking – for each $\mathscr{X}$-family – convex combinations of the predictive (count) previsions from $\sigma^{\max}$ and $\sigma^{\min}$ for $\check{n}:1..N-1$: (let f

be a gamble on $\mathscr{X}$ and $\check{m}$ an observed count vector from $\check{n}^{\mathscr{X}}$)

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) := \varepsilon_{\check{n}} \cdot \mathrm{Wa}(f|\check{m}) + (1-\varepsilon_{\check{n}}) \cdot \min\{f\}. \tag{3.95}$$

Of course the initial predictive (count) prevision is the vacuous prevision; i.e., $\underline{P}_{\mathscr{X}} f = \underline{Q}_{\mathscr{X}} f = \min\{f\}$. This is a convex mixture of the initial predictive previsions in $\sigma^{\max}$ and $\sigma^{\min}$ in a very trivial sense.

The lower probability function q^{ε} corresponding to a mixing predictive system σ^{ε} is the convex combination $k,\ell : (\mathbb{N}_{<N})^2 \wedge \ell \leq k > 0 ; \varepsilon_k \cdot \frac{\ell}{k}$ of the lower probability functions $q^{\max}$ and $q^{\min}$ of the vacuous and Haldane predictive systems. We see that it is necessary for representation insensitivity that ε be categorization-independent. This is also sufficient, as representation insensitivity is preserved by taking convex mixtures of any representation insensitive predictive system.

Let $k : 1..N-1$, then $\varepsilon_k = q^{\varepsilon}(k,k)$ and $1-\varepsilon_k = 1-q^{\varepsilon}(k,k)$. So ε_k *is the lower probability of observing a nontrivial event that has always been observed before* and $1-\varepsilon_k$ *is the upper probability of observing a nontrivial event that has never been observed before.* The latter characterizes the imprecision of the predictive prevision; so the imprecision depends on the number of observations directly through the mixing factor ε_k. The lower probability function of a mixing system is additive in its first argument; e.g., let $\ell : \mathbb{N}_{\leq k}$, then the lower probability $q^{\varepsilon}(k,\ell)$ of observing a nontrivial event ℓ times out of k trials is ℓ times the lower probability $q^{\varepsilon}(k,1) = \varepsilon_k \cdot \frac{1}{k}$ of observing that event once.

The investigation to see under which conditions a mixing predictive system can be exchangeable will take place in a number of more technical stages. Some readers may wish to skip ahead to the resulting expression $(3.101)_{135}$ for the predictive previsions that replaces (3.95) – which ensures that exchangeable joints exist – and then continue with the first paragraph after that.

The first stage consists of drawing an initial conclusion from the necessary condition for exchangeability $(3.87)_{127}$ we derived before. Translated in terms of mixing sequence components, this condition becomes: (additionally assume $k < N-1$)

$$\begin{aligned} &\tfrac{\ell}{k} \cdot \varepsilon_k \geq \tfrac{\ell}{k+1} \cdot \varepsilon_{k+1} + \tfrac{\ell}{k} \cdot \varepsilon_k \cdot (\tfrac{\ell+1}{k+1} \cdot \varepsilon_{k+1} - \tfrac{\ell}{k+1} \cdot \varepsilon_{k+1}), \\ \text{or} \qquad &(k+1) \cdot \varepsilon_k \geq (k+\varepsilon_k) \cdot \varepsilon_{k+1}. \end{aligned} \tag{3.96}$$

So if $\varepsilon_k = 0$, then ε_{k+1} must also be zero for the system to be exchangeable. This means that if we take $\varepsilon_1 = 0$, then the predictive system must coincide with the vacuous predictive system. In other words: *if we want to learn, we must start doing so from the very beginning.*

We do want to learn, so from now onwards, we assume $\varepsilon :]0,1]^{N-1}$. This assumption is useful in our second stage: we now replace the expression for the vacuous prevision in (3.95) by its expression as a lower

envelope (cf. $(1.55)_{50}$):

$$\begin{aligned}\underline{Q}_{\mathscr{X}}(f|\check{m}) &= \varepsilon_{\check{n}} \cdot \mathrm{Wa}(f|\check{m}) + (1-\varepsilon_{\check{n}}) \cdot \min_{z:\mathscr{X}} \mathrm{Wa}(f|C_{\mathscr{X}} z) \\ &= \min_{z:\mathscr{X}} \mathrm{Wa}\big(f \,\big|\, \varepsilon_{\check{n}} \cdot \tfrac{\check{m}}{\check{n}} + (1-\varepsilon_{\check{n}}) \cdot C_{\mathscr{X}} z\big)\end{aligned}$$

where we used the weighted average's definition $(3.90)_{128}$ twice. To write the weights more elegantly, divide the numerator and denominator by $\frac{\varepsilon_{\check{n}}}{\check{n}}$ and define the nonnegative real number $s_{\check{n}} := \check{n} \cdot \frac{1-\varepsilon_{\check{n}}}{\varepsilon_{\check{n}}}$ to obtain

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) = \min_{z:\mathscr{X}} \mathrm{Wa}(f|\check{m} + s_{\check{n}} \cdot C_{\mathscr{X}} z) = \min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(f|\check{m} + s_{\check{n}} \cdot t),$$

where in the second, equivalent expression, the minimization ranges over the whole simplex and thus over the whole credal set, not only over its extreme points (cf. $(1.55)_{50}$). Condition (3.96), translated, requires that $s_{k+1} \geq s_k$ for all k in $1..N-2$.

Now, to investigate the exchangeability of predictive systems consisting of predictive previsions of this type, we must look at the joint (updated) sequence distributions defined by them. For this, we need to take the same steps as those taken for proving the exchangeability of the Haldane predictive system in the previous subsection.

This brings us to the third stage, where we start by generalizing the expression $(3.92)_{129}$ for the situation where $\check{n} := N-2$. We know that marginal extension $(1.84)_{62}$ results in the least committal coherent joint, but that this joint is not necessarily the least committal coherent and *exchangeable* joint. Let f be a gamble on $\mathscr{X}^2$, then $\underline{P}^2_{\mathscr{X}}(f|\check{x})$ must not be smaller than the (lower envelope version $(1.86)_{62}$ of) marginal extension

$$\min_{t:\Delta_{\mathscr{X}}} \min_{r:(\Delta_{\mathscr{X}})^{\mathscr{X}}} \mathrm{Wa}\big(\mathrm{Wa}(f|\check{m} + C_{\mathscr{X}} \bullet + s_{N-1} \cdot r_\bullet) \,\big|\, \check{m} + s_{N-2} \cdot t\big). \quad (3.97)$$

To be more precise: as $\underline{P}^2_{\mathscr{X}}(\bullet|\check{x})$ is exchangeable if and only if its credal set consists of exchangeable linear previsions only (cf. $(3.6)_{99}$), this credal set must be equal to or a subset of the set of exchangeable linear previsions in the credal set of the marginal extension.

A linear prevision in the credal set of this marginal extension is characterized by the frequency vectors t and r_z ($z : \mathscr{X}$) from $\Delta_{\mathscr{X}}$: (cf. $(3.92)_{129}$)

$$\begin{aligned}&\mathrm{Wa}\big(\mathrm{Wa}(f|\check{m} + C_{\mathscr{X}} \bullet + s_{N-1} \cdot r_\bullet) \,\big|\, \check{m} + s_{N-2} \cdot t\big) \\ &\quad = \textstyle\sum_{\hat{y}:\mathscr{X}^2} f\hat{y} \cdot \dfrac{(\check{m} + C_{\mathscr{X}} \hat{y}_1 + s_{N-1} \cdot r_{\hat{y}_1})_{\hat{y}_2}}{s_{N-1} + N - 1} \cdot \dfrac{(\check{m} + s_{N-2} \cdot t)_{\hat{y}_1}}{s_{N-2} + N - 2} \\ &\quad = \textstyle\sum_{\hat{y}:\mathscr{X}^2} f\hat{y} \cdot \dfrac{\big((\check{m} + s_{N-1} \cdot r_{\hat{y}_1})_{\hat{y}_2} + \delta_{\hat{y}_1 \hat{y}_2}\big) \cdot (\check{m} + s_{N-2} \cdot t)_{\hat{y}_1}}{(s_{N-1} + N - 1) \cdot (s_{N-2} + N - 2)}. \quad (3.98)\end{aligned}$$

To check that convex combinations can always be rewritten in this form, let $\lambda : [0,1]$, $t', t'' : (\Delta_{\mathscr{X}})^2$, and $t := \lambda \cdot t' + (1-\lambda) \cdot t''$; for all $z : \mathscr{X}_t$, let $r'_z, r''_z : (\Delta_{\mathscr{X}})^2$ and $r_z := \lambda \cdot \frac{t'_z}{t_z} \cdot r'_z + (1-\lambda) \cdot \frac{t''_z}{t_z} \cdot r''_z$. Then

$$\lambda \cdot \mathrm{Wa}\big(\mathrm{Wa}(f|r'_\bullet) \,\big|\, t'\big) + (1-\lambda) \cdot \mathrm{Wa}\big(\mathrm{Wa}(f|r''_\bullet) \,\big|\, t''\big) = \mathrm{Wa}\big(\mathrm{Wa}(f|r_\bullet) \,\big|\, t\big).$$

We see that the prevision of $(3.98)_\frown$ is exchangeable when the second factor's numerator is invariant under permutation of the sequence order; i.e., if for all nonidentical $\hat{y}_1$ and $\hat{y}_2$ in $\mathscr{X}$

$$\frac{\check{m}_{\hat{y}_2}}{s_{N-2}} \cdot r_{\hat{y}_2\hat{y}_1} + \frac{\check{m}_{\hat{y}_1}}{s_{N-1}} \cdot t_{\hat{y}_2} + t_{\hat{y}_2} \cdot r_{\hat{y}_2\hat{y}_1} = \frac{\check{m}_{\hat{y}_1}}{s_{N-2}} \cdot r_{\hat{y}_1\hat{y}_2} + \frac{\check{m}_{\hat{y}_2}}{s_{N-1}} \cdot t_{\hat{y}_1} + t_{\hat{y}_1} \cdot r_{\hat{y}_1\hat{y}_2}. \tag{3.99}$$

In the very specific, simple case where $N = 2$ and $\mathscr{X} = \{a, b\}$, $\check{m} = 0$ and we find that the exchangeability condition for joint linear previsions above reduces to $t_b \cdot r_{ba} = t_a \cdot r_{ab}$ (always satisfied for $r_{ba} = 0 = r_{ab}$), or, because $t_b = 1 - t_a$, to $t_a = \frac{r_{ba}}{r_{ba}+r_{ab}}$ and $t_b = \frac{r_{ab}}{r_{ba}+r_{ab}}$. So for this case, the least committal and exchangeable joint lower prevision $\underline{P}^2_{\{a,b\}}$ becomes (cfr. $(3.97)_\frown$; f is still a gamble on $\mathscr{X}^2$ and $g := z : \mathscr{X} \,;\, f(z,z)$)

$$\min\Big\{\min_{t:\Delta_{\{a,b\}}} \mathrm{Wa}(g|t),\ \min_{r:(\Delta_{\{a,b\}})^{\{a,b\}} \wedge r_{ba}+r_{ab}>0} \mathrm{Wa}\big(\mathrm{Wa}(f|C_{\mathscr{X}}\bullet + s_1 \cdot r_\bullet) \,\big|\, (r_{ba}, r_{ab})\big)\Big\}.$$

This expression is already rather complex, given the simplicity of this case. It forebodes the complexities one encounters when trying to derive expressions for the least committal and exchangeable joint of more than two marginals when $N > 2$ and $|\mathscr{X}| > 2$. It would be very interesting to obtain these expressions… as a future challenge.

In this thesis, we now work towards other interesting expressions, which correspond to a simpler, more restricted case: Of the exchangeable linear previsions dominating $(3.97)_\frown$, we only consider those for which the so-called hyperparameters r_z and t are equal for all z in $\mathscr{X}$; we do this for all concatenations of successive marginal and conditional previsions. We call this restriction made out of mathematical convenience the 'constant hyperparameter path'. From (3.99), it then follows that $s_{N-1} = s_{N-2}$.

Assume $s := s_{N-1} = s_{N-2} > 0$ (otherwise we would have a Haldane joint), then we now know that

$$\underline{P}^2_{\mathscr{X}}(f|\check{x}) := \min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}\big(\mathrm{Wa}(f|\check{m} + C_{\mathscr{X}}\bullet + s \cdot t) \,\big|\, \check{m} + s \cdot t\big)$$

is the least committal exchangeable constant hyperparameter path joint. We can use a minimum over the compact unit simplex, because the concatenation of two weighted averages results in an expression for a linear prevision that is a (continuous) polynomial function in t (see $(3.98)_\frown$); this generalizes to the a similar concatenation of an arbitrary number of weighted averages. The $\mathscr{X}$-marginals of $\underline{P}^2_{\mathscr{X}}(f|\check{x})$ correspond to the predictive prevision for X_{N-1} (to understand why they are identical, see the next-to-last paragraph before §3.1.2$_{99}$). Also, $\min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(\bullet|\check{m}+C_{\mathscr{X}}z+s\cdot t)$ is a jointly coherent updated prevision after observing X_{N-1} to be $z : \mathscr{X}$,

even when $\min_{t:\Delta_{\mathscr{X}}}(\check{m}_z + s\cdot t_z) = 0$, because then we can use regular extension $(1.70)_{58}$.

The reasoning leading to the expression for this joint started out by taking the marginal extension of the predictive previsions for the random variables X_{N-1} and X_N. Our fourth stage is based on the fact that it equally applies when using the marginal extension for all pairs of random variables X_k and X_{k+1}, where $k:2..N-1$. Therefore, to ensure exchangeability when walking the constant hyperparameter path, $s_{k-1} = s_k$ must hold for all k, or, in other words, the sequence $k:1..N-1\,;\, k\cdot 1-\varepsilon_k/\varepsilon_k$ must be constant. If some s in $\mathbb{R}_{>0}$ is fixed, we get $\varepsilon := k:1..N-1\,;\, k/s+k$ and the expression $(3.95)_{132}$ for the corresponding mixing predictive system becomes (let f be a gamble on $\mathscr{X}$ and thus, by isomorphism, on $1^{\mathscr{X}}$)

$$\underline{P}_{\mathscr{X}}(f|\check{m}) = \underline{Q}_{\mathscr{X}}(f|\check{m}), \tag{3.100}$$

$$\underline{Q}_{\mathscr{X}}(f|\check{m}) = \underline{\mathrm{Dm}}(f|1,\check{m},s) := \underline{\mathrm{Wa}}(f|\check{m},s), \tag{3.101}$$

where in the last expression we used

$$\underline{\mathrm{Wa}}(f|\check{m},s) := \min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(f|\check{m}+s\cdot t) \tag{3.102}$$

$$= \tfrac{\check{n}}{\check{n}+s}\cdot \mathrm{Wa}(f|\check{m}) + \tfrac{s}{\check{n}+s}\cdot \min\{f\}. \tag{3.103}$$

In this expression, s takes a similar role as $\check{n}$, the total number of observed counts; therefore s is often referred to as 'the number of pseudocounts'. (The reason for introducing the notation $\underline{\mathrm{Dm}}(\bullet|\bullet,\bullet,s)$ will become clear in the next paragraph.)

So now the allowed concatenations of the linear previsions in the credal sets of the predictive previsions and parameters involved are constrained in such a way that there is only one possible updated joint prevision for each sequence of observations $\check{x}$. It is obtained – in this fifth stage – by retracing the steps leading to the corresponding updated joint prevision $(3.94)_{130}$ for the Haldane system (concatenating weighted averages), but substituting $\check{m}$ with $\check{m}+s\cdot t$ and minimizing over all t in $\Delta_{\mathscr{X}}$: (let f be a gamble on $\mathscr{X}^{\hat{n}}$ and h a gamble on $\hat{n}^{\mathscr{X}}$)

$$\underline{P}^{\hat{n}}_{\mathscr{X}}(f|\check{x}) = \underline{Q}^{\hat{n}}_{\mathscr{X}}\big(\mathrm{Mh}(f|\bullet)\,\big|\, C_{\mathscr{X}}\check{x}\big), \tag{3.104}$$

$$\underline{Q}^{\hat{n}}_{\mathscr{X}}(h|\check{m}) = \underline{\mathrm{Dm}}(h|\hat{n},\check{m},s) := \min_{t:\Delta_{\mathscr{X}}} \mathrm{Dm}(h|\hat{n},\check{m}+s\cdot t), \tag{3.105}$$

where we have used the Dirichlet-multinomial linear prevision that corresponds to the Dirichlet-compound multinomial distribution. Let k be some positive integer and $\alpha:(\mathbb{R}^{\mathscr{X}})_{\geq 0\wedge\neq 0}$, then this prevision is defined for any gamble g on $k^{\mathscr{X}}$ by [Johnson et al. 1997, §35.13.1_{80}]

$$\mathrm{Dm}(g|k,\alpha) := \frac{1}{\binom{\sum\alpha+k-1}{k}}\cdot \sum_{\hat{m}:k^{\mathscr{X}_\alpha}} g\hat{m}\cdot \prod_{z:\mathscr{X}_\alpha}\binom{\alpha_z+\hat{m}_z-1}{\hat{m}_z} \tag{3.106}$$

$$= \mathrm{Dm}\big(g\circ(\bullet_{\mathscr{X}})\,\big|\, k,\alpha_{\mathscr{X}_\alpha}\big), \tag{3.107}$$

Generalized binomial coefficients: (let $\ell:\mathbb{N}$ and $\beta:\mathbb{R}_{>\ell-1}$)

$$\binom{\beta}{\ell} := \frac{\prod_{i:0..\ell-1}(\beta-i)}{\ell!} = \frac{\Gamma(\beta+1)}{\ell!\cdot\Gamma(\beta-\ell+1)},$$

where Γ denotes the gamma function [Abramowitz & Stegun 1972, §6_{253}].

where the second line compactly shows (using trivial extension) the effect of having components of α that are zero.

For our sixth and final stage, we can rewrite the expression for the initial predictive prevision as follows: (let f be a gamble on $\mathscr{X}$)

$$\underline{Q}_{\mathscr{X}} f = \min\{f\} = \min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(f|t) = \min_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(f|s\cdot t),$$

which has the same form $(3.101)_{\frown}$ as all the other mixing system predictive previsions. So in contrast to the calculation of the Haldane prior, the prior has the same form as the posterior joints: (now let f be a gamble on $\mathscr{X}^N$ and h a gamble on $N^{\mathscr{X}}$)

$$\underline{P}^N_{\mathscr{X}} f = \underline{Q}^N_{\mathscr{X}}\big(\mathrm{Mh}(f|\circ)\big), \tag{3.108}$$

$$\underline{Q}^N_{\mathscr{X}} h = \underline{\mathrm{Dm}}\big(h \,\big|\, N, (\mathscr{X};0), s\big) := \min_{t:\Delta_{\mathscr{X}}} \mathrm{Dm}(h|N, s\cdot t), \tag{3.109}$$

So what we ended up with here is the predictive system characterized by $(3.101)_{\frown}$, augmented by its respectively prior and posterior joints (3.108)–(3.109) and (3.104)–$(3.105)_{\frown}$. Together, these form Walley & Bernard's [1999] imprecise Dirichlet-multinomial model or IDMM. In some sense, we justified this model by constructing it from first principles:

(i) requiring that its predictive previsions form a representation insensitive system,
(ii) requiring that this system is exchangeable, i.e., find exchangeable joints,
(iii) restricting ourselves to linear-vacuous predictive previsions out of mathematical convenience, and
(iv) walking the constant hyperparameter path out of mathematical convenience.

Walley [1991, §2.9.2$_{93}$] proposes a behavioral interpretation for linear-vacuous mixtures: a fractional tax on any gain above the minimum possible one. For the IDMM, the rate is $\frac{s}{s+\tilde{n}}$, which in a sense punishes speculation.

There is *only one* free parameter in this inference model that has to be fixed before using it: a positive real number of pseudocounts which determines the speed of learning; low values beget fast, or even rash learners, higher values beget slower, more conservative learners. I have not yet heard any convincing argument to prefer one pseudocount value over another a priori. My current attitude is that the choice depends on the specifics of the application the model is used for; e.g., for artificial intelligence applications in user interfaces, it could be determined with a usability test. Walley [1996] and Walley & Bernard [1999] mainly work with $s := 1$ or $s := 2$.

With our rediscovery of the IDMM, we arrived at the most important result of this section. The IDMM will also be used to bootstrap the next section about parametric categorical inference. But before we do that, we show that the IDMM predictive system σ^s – i.e., the one defined by $(3.101)_{\frown}$ – satisfies another very interesting property besides exchangeability and representation insensitivity.

3.2.6 *Specificity*

We now consider the situation where, in addition to the standard assumptions of having observed some sample $\check{m}$, we know that the next observation belongs to a proper subset A of the category set $\mathscr{X}$. This could for example be the case when the actual observation has been made, but it was imperfect. We now ask what our (immediate) predictive model is, when starting from some mixing system σ^{ε} characterized by some mixing sequence $\varepsilon : [0,1]^{N-1}$, and when taking into account the information gained by the imperfect observation.

To answer this question, we condition the predictive count prevision $\underline{Q}_{\mathscr{X}}(\bullet|\check{m})$ on A using the GBR. We know that when $\underline{Q}_{\mathscr{X}}(A|\check{m}) = \varepsilon_{\check{n}} \cdot \sum \check{m}_A/\check{n} = 0$, the updated predictive prevision $\underline{Q}_{\mathscr{X}}(\bullet|\check{m}, A)$ will be vacuous; this can happen when $\varepsilon_{\check{n}} = 0$ or $\sum \check{m}_A = 0$. Whenever $\varepsilon_{\check{n}} \cdot \sum \check{m}_A/\check{n} > 0$, we can obtain the updated predictive prevision by using the functional form of the GBR $(1.82)_{61}$: (let f be a gamble on A and μ the tentative root, some real number)

$$\begin{aligned}
0 &= \underline{Q}_{\mathscr{X}}\big((f-\mu)_{\mathscr{X}} \,\big|\, \check{m}\big) \\
&= \varepsilon_{\check{n}} \cdot \mathrm{Wa}\big((f-\mu)_{\mathscr{X}} \,\big|\, \check{m}\big) + (1-\varepsilon_{\check{n}}) \cdot \min\big\{(f-\mu)_{\mathscr{X}}\big\} \\
&= \varepsilon_{\check{n}} \cdot \tfrac{\sum \check{m}_A}{\check{n}} \cdot \mathrm{Wa}(f-\mu|\check{m}_A) + (1-\varepsilon_{\check{n}}) \cdot \min\big\{0, \min\{f-\mu\}\big\} \\
&= \varepsilon_{\check{n}} \cdot \tfrac{\sum \check{m}_A}{\check{n}} \cdot \big(\mathrm{Wa}(f|\check{m}_A) - \mu\big) + (1-\varepsilon_{\check{n}}) \cdot \big(\min\big\{\mu, \min\{f\}\big\} - \mu\big)
\end{aligned}$$

The right-hand side can only be zero when $\mu \geq \min\{f\}$, because by definition $\min\{f\} \leq \mathrm{Wa}(f|\check{m}_A)$. This observation allows us to find the solution

$$\begin{aligned}
\underline{Q}_{\mathscr{X}}(f|\check{m}, A) &= \mu \\
&= \tfrac{\varepsilon_{\check{n}} \cdot \sum \check{m}_A/\check{n}}{\varepsilon_{\check{n}} \cdot \sum \check{m}_A/\check{n} + 1 - \varepsilon_{\check{n}}} \cdot \mathrm{Wa}(f|\check{m}_A) + \tfrac{1-\varepsilon_{\check{n}}}{\varepsilon_{\check{n}} \cdot \sum \check{m}_A/\check{n} + 1 - \varepsilon_{\check{n}}} \cdot \min\{f\} \\
&= \tfrac{\sum \check{m}_A}{\sum \check{m}_A + \check{n} \cdot (1-\varepsilon_{\check{n}})/\varepsilon_{\check{n}}} \cdot \mathrm{Wa}(f|\check{m}_A) + \tfrac{\check{n} \cdot (1-\varepsilon_{\check{n}})/\varepsilon_{\check{n}}}{\sum \check{m}_A + \check{n} \cdot (1-\varepsilon_{\check{n}})/\varepsilon_{\check{n}}} \cdot \min\{f\} \\
&= \frac{\check{n}_A}{\check{n}_A + s_{\check{n}}} \cdot \mathrm{Wa}(f|\check{m}_A) + \frac{s_{\check{n}}}{\check{n}_A + s_{\check{n}}} \cdot \min\{f\},
\end{aligned}$$

where in the last line, we have written $s_{\check{n}} := \check{n} \cdot 1-\varepsilon_{\check{n}}/\varepsilon_{\check{n}} > 0$ and $\check{n}_A := \sum \check{m}_A$ for convenience. Notice that it also subsumes the case $\check{n}_A = \sum \check{m}_A = 0$, as well as the case $\varepsilon_{\check{n}} = 0$ by taking the limit for $s_{\check{n}}$ going to ∞.

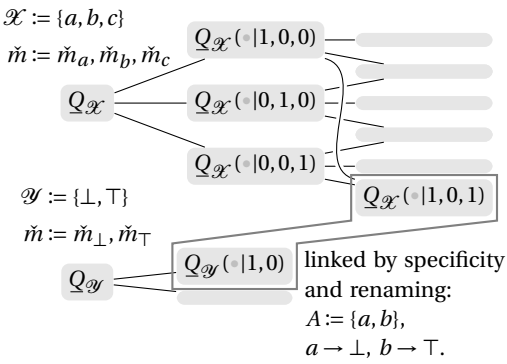
It is interesting to compare this expression to

$$\underline{Q}_A(f|\check{m}_A) = \frac{\check{n}_A}{\check{n}_A + s_{\check{n}_A}} \cdot \mathrm{Wa}(f|\check{m}_A) + \frac{s_{\check{n}_A}}{\check{n}_A + s_{\check{n}_A}} \cdot \min\{f\},$$

where $s_{\check{n}_A} := \check{n}_A \cdot \frac{1-\varepsilon_{\check{n}_A}}{\varepsilon_{\check{n}_A}}$. They are equal whenever

$$\frac{\check{n}_A \cdot (s_{\check{n}} - s_{\check{n}_A})}{(\check{n}_A + s_{\check{n}}) \cdot (\check{n}_A + s_{\check{n}_A})} \cdot \big(\mathrm{Wa}(f|\check{m}_A) - \min\{f\}\big) = 0.$$

If we want this equality $\underline{Q}_{\mathscr{X}}(f|\check{m}, A) = \underline{Q}_A(f|\check{m}_A)$ to hold for all catego-



rizations $\mathscr{X}$, for all count vectors $\check{m}$, all events A, and all gambles f, then it must hold that $s_{\check{n}} = s_{\check{n}_A}$ for every $\check{n}$. To see this, consider some nontrivial case, e.g., $\mathscr{X} := \{a, b, c\}$, $(\check{m}_a, \check{m}_b, \check{m}_c) := (\check{n} - 1, 0, 1)$, $A := \{a, b\}$, and $f := I^a$.)

The above equality is called the specificity property and any predictive system satisfying it is called specific. This terminology was first coined by Bernard [1997]. What we have found in the previous paragraphs is that IDMM predictive systems are the only specific mixing predictive systems.

From a mathematical point of view, specificity already seems like a nice property, but when we reflect on its meaning, we see that in some contexts it is an appropriate and intuitively appealing property for a predictive system to possess: It ensures that the inferences for the next observation, when taking into account that that next observation belongs to some known set of categories, only depend on the relative number of observations of categories within the known set. For example, when suggesting a book to a person, it will not help you to know how many servings of which legumes that person has eaten throughout her life.

That specificity is not always a desirable property must also be clear: when making guesses about which predator is observed next in some ecosystem, it might not be prudent to neglect the relative number of predator-specific prey that have been observed in the area. Moreover, in such a context with meaningful links between the categories, although requiring exchangeability may still seem like a good idea, requiring representation insensitivity is not. (This remark is just meant to stress once again that the models we are working with here are not universally valid.)

3.3 PARAMETRIC INFERENCE: THE IMPRECISE DIRICHLET MODEL

This section starts with a natural variation on the context of the previous one: we consider a subject who is making an arbitrary length sequence of observations of a certain phenomenon. We assume the sequence to be infinitely exchangeable. In §3.1.3$_{103}$ and §3.1.5$_{112}$ we have respectively learned how to model our uncertainty about such sequences – using a frequency distribution – and how to update this model.

We start this short section by *deriving* the imprecise Dirichlet model (IDM), a parametric inference model proposed by Walley [1996]. The second and last subsection discusses conjugate updating and how it relates to what we have discovered.

3.3.1 *From the* IDMM *to the* IDM *via infinite exchangeability*

At the end of §3.2.5$_{131}$, we discovered the IDMM, a predictive inference model for categorical data – a sequence of N samples – described by a coherent set of count distributions $\underline{\mathrm{Dm}}(\bullet|\hat{n},\check{m},s)$, where s is some positive real number of pseudocounts, $\check{m}$ is some count vector of size $\sum\check{m}=\check{n}$ (with $\check{n}:0..N-1$) and $\hat{n}:1..N-\check{n}$ is the number of observations we are making predictions about. This set consists of both marginal previsions (e.g., cf. (3.101)$_{135}$ for $\check{n}=1$) and prior and posterior exchangeable joint previsions (3.109)$_{136}$ and (3.105)$_{135}$.

Now, if we let N become arbitrarily large, our uncertainty model will describe the sample sequence as being infinitely exchangeable if $\infty\text{-cns}\{\underline{\mathrm{Dm}}(\bullet|\hat{n},\check{m},s)\mid\hat{n}:\mathbb{N}_{>0}\}$ for all observed count vectors $\check{m}$ of all sizes $\check{n}:\mathbb{N}$. In other words, following (3.29)$_{106}$, if we can find a coherent frequency distribution $\underline{R}$ on $\mathscr{C}_{\Delta_{\mathscr{X}}}$, or, equivalently, on $\mathscr{V}_{\Delta_{\mathscr{X}}}$, for every $\check{m}$ such that $\underline{\mathrm{Dm}}(\bullet|\hat{n},\check{m},s)=\underline{R}\big(\mathrm{Cm}(\bullet|\hat{n},\bullet)\big)$.

It is not very hard to find such a frequency distribution, so the answer is yes. The name of the type of linear prevision that defines the IDMM, the Dirichlet-multinomial prevision, originates from the fact that the Dirichlet-multinomial distribution can be seen as a compounding of a multinomial with a Dirichlet distribution [Johnson et al. 1997, §35.13.1$_{80}$]; to wit, the Dirichlet distribution is used as a second-order model, to express uncertainty about the relative frequencies used in the multinomial distribution (i.e., the frequency vector $\vartheta:\Delta_{\mathscr{X}}$ in (3.26)$_{104}$ or (3.33)$_{107}$). Therefore every extreme point $\mathrm{Dm}(\bullet|\hat{n},\check{m}+s\cdot t)$ of $\mathscr{M}\big(\underline{\mathrm{Dm}}(\bullet|\hat{n},\check{m},s)\big)$ (where $t:\Delta_{\mathscr{X}}$) can be written as $\mathrm{Di}\big(\mathrm{Cm}(\bullet|\hat{n},\bullet)\mid\check{m}+s\cdot t\big)$, where we have used the Dirichlet linear prevision that corresponds to the Dirichlet distribution [Kotz et al. 2000, §49.1$_{485}$]. Let α be a nonnegative nonzero element of $\mathbb{R}^{\mathscr{X}}$, then this prevision is defined for any h in $\mathscr{C}_{\Delta_{\mathscr{X}}}$ by

$$\mathrm{Di}(h|\alpha):=\frac{\Gamma(\sum\alpha)}{\prod_{z:\mathscr{X}_\alpha}\Gamma\alpha_z}\cdot\int_{\Delta_{\mathscr{X}_\alpha}}h\vartheta_{\mathscr{X}}\cdot\Big(\prod_{z:\mathscr{X}_\alpha}\vartheta_z^{\alpha_z-1}\Big)\,\mathrm{d}\vartheta \tag{3.110}$$

$$=\mathrm{Di}\big(h\circ(\bullet_{\mathscr{X}})\mid\alpha_{\mathscr{X}_\alpha}\big), \tag{3.111}$$

where the second line compactly shows (using trivial extension) the effect of having components of α that are zero. Johnson et al. [1997, §35.13.1$_{80}$] present the compounding in terms of probability density and mass functions parameterized by a pointwise strictly positive α; to make the correspondence with our expressions fit, use (3.107)$_{135}$, (3.111), and the fact that $\mathrm{Cm}(g|k,\vartheta_{\mathscr{X}})=\mathrm{Cm}\big(g\circ(\bullet_{\mathscr{X}})\mid k,\vartheta\big)$ for any $k:\mathbb{N}_{>0}$, $\vartheta:k^{\mathscr{X}_\alpha}/k$, and any gamble g on $k^{\mathscr{X}}$.

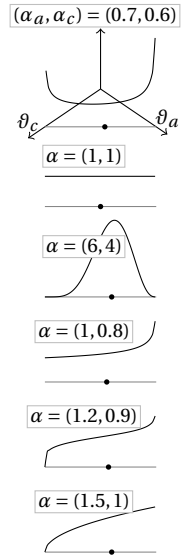
So then

$$\underline{\mathrm{Dm}}(\bullet|\hat{n},\check{m},s)=\underline{\mathrm{Di}}\big(\mathrm{Cm}(\bullet|\hat{n},\bullet)\mid\check{m},s\big), \tag{3.112}$$

where

$$\underline{\mathrm{Di}}(\bullet|\check{m},s):=\min_{t:\Delta_{\mathscr{X}}}\mathrm{Di}(\bullet|\check{m}+s\cdot t), \tag{3.113}$$

Some Dirichlet densities on $\Delta_{\{a,c\}}$: ($\alpha/\sum\alpha$ indicated with a dot)



Notice how the border behavior depends on the sign of the components of $\alpha-1$; the same is seen for higher dimensions.

is the IDM-posterior for the observed count vector $\check{m}$ and pseudocounts s. When $\check{m} = 0$, we obtain the IDM-prior

$$\underline{\mathrm{Di}}\big(\bullet \,\big|\, (\mathscr{X}; 0), s\big) := \min_{t:\Delta_{\mathscr{X}}} \mathrm{Di}(\bullet|s \cdot t). \tag{3.114}$$

Note that we have only defined the IDM $\underline{\mathrm{Di}}(\bullet|\bullet, s)$ for *continuous* gambles on $\Delta_{\mathscr{X}}$ (cf. last paragraph before §3.1.4$_{107}$). We do not investigate the interesting question of extension to all (measurable) gambles.

With (3.114) and (3.113)$_{\cap}$, we have respectively found the coherent prior and posterior frequency distributions that constitute Walley's [1996] imprecise Dirichlet model or IDM. In some sense we justified it by deriving it from first principles. These frequency distributions

(i) model the uncertainty about a sample sequence that is assumed to be infinitely exchangeable,
(ii) characterize exchangeable representation insensitive mixing predictive systems (with a mixing sequence determined by s) for arbitrary sequence lengths, and
(iii) are the least committal (most conservative) previsions satisfying these requirements when walking the constant hyperparameter path.

3.3.2 *Conjugate updating*

We have derived the IDM (via the IDMM) by starting out with a family of predictive previsions and using these marginals to build prior and posterior joints. As seen in §3.1.6$_{115}$, in classical Bayesian updating one usually approaches the updating problem differently: a prior joint distribution is proposed and using Bayes's rule this is updated with a likelihood function to a posterior joint distribution.

Seen in such a context, the prior IDM (3.114) is only one among many possible priors. Because of the way we derived it, we know that it is special and unique in many ways (and I have actually become quite attached to it). To its further distinction, the IDM has another interesting property: it is conjugate to the multinomial likelihood, or closed under categorical sampling with replacement. Roughly speaking, this means that both the prior and posterior uncertainty models have the same functional form [Bernardo & Smith 1994, §5.2$_{265}$]; compare (3.114) with (3.113)$_{\cap}$ to see that this is indeed the case. This conjugacy property has the advantage of greatly increasing the mathematical tractability of the updating process; we know of no behavioral justification for requiring this property.

Conjugacy in an updating context is totally unrelated to the conjugacy of lower and upper previsions (1.14)$_{36}$.

However, we *do* know at which point in the construction of the IDM it acquired this property. Using the notation of the previous subsection, recall that the immediate predictive previsions corresponding to an IDM $\underline{\mathrm{Di}}(\bullet|\check{m}, s)$ are linear-vacuous previsions: (let f be a gamble on $\mathscr{X}$, and

use $(3.112)_{139}$, $(3.31)_{106}$, $(3.101)_{135}$, and $(3.103)_{135}$)

$$\underline{\mathrm{Di}}\big(\mathrm{Mn}(f|1,\circledast)\,\big|\,\check{m},s\big) = \underline{\mathrm{Dm}}\big(\mathrm{Mh}(f|\circledast)\,\big|\,1,\check{m},s\big) = \underline{\mathrm{Wa}}(f|\check{m},s) \qquad (3.115)$$
$$= \tfrac{\check{n}}{s+\check{n}}\cdot\mathrm{Wa}(f|\check{m}) + \tfrac{s}{s+\check{n}}\cdot\min\{f\}.$$

Or, in other words, that each of the precise Dirichlet models $\mathrm{Di}(\circledast|\check{m}+s\cdot t)$ – where $t:\Delta_{\mathscr{X}}$ – that defines this IDM (cf. $(3.113)_{139}$) is a weighted average:

$$\mathrm{Di}(f|\check{m}+s\cdot t) = \tfrac{\check{n}}{s+\check{n}}\cdot\mathrm{Wa}(f|\check{m}) + \tfrac{s}{s+\check{n}}\cdot\mathrm{Wa}(f|t) = \mathrm{Wa}(f|\check{m}+s\cdot t).$$

This is a property that *characterizes* conjugate families of distributions for many likelihoods, among them the multinomial one [Diaconis & Ylvisaker 1979, just above the Acknowledgment].

It is therefore no surprise that we find no behavioral justification for conjugate updating: it is a consequence of our choice to restrict ourselves to mixing predictive systems out of mathematical convenience (cf. §3.2.5$_{131}$). Linear-vacuous mixtures – also called contamination models or neighborhood models – are typically used in so-called robust Bayesian analysis. Therefore the IDM and other neighborhood-based uncertainty models are better supported by the so-called sensitivity analysis interpretation of imprecise probabilities [Walley 1991, §1.1.5$_{6}$]: There is an ideal, *precise* uncertainty model, but due to practical limitations it cannot be known precisely; so to deal with our limitations honestly, we should use imprecise-probabilistic models (typically neighborhood models and parameterized models such as the IDM).

Berger et al. [1994] give an overview of robust Bayesian analysis.

One could also say that the IDMM is conjugate to the multivariate hypergeometric likelihood, or closed under categorical sampling without replacement, but this is not typical usage. A reason is that the posterior lower previsions of the IDMM $(3.105)_{135}$ are defined on a different (smaller) possibility space than the corresponding prior $(3.109)_{136}$. This also corresponds to the fact that the multivariate hypergeometric likelihood functions' domain varies with the number of already observed samples (cf. (3.38)–$(3.41)_{109}$).

Normally only parametric inference models – and their defining previsions or probability mass or density functions – are said to be conjugate to some likelihood. Actually, there is a class of likelihood functions for which a conjugate inference model exists that shares many properties with the one member we have already encountered, the multinomial likelihood. The most important of these properties are that

(i) they are all determined by a finite dimensional sufficient statistic,
(ii) and they share a common functional form.

This is the class of the so-called regular exponential families. It turns out that the analogies between these families and the multinomial one are so strong that many aspects of what we have seen in this chapter about learning from categorical samples can be generalized to the other

sampling models in the class. This is the subject of the next chapter, 'Inference models for exponential families'$_{154}$.

The rest of this chapter is devoted to some applications of the inference models we have constructed from first principles.

3.4 APPLICATIONS

In this section, we apply the IDM, i.e., what we have learned up until now in this chapter, to two slightly more concrete problems. The first problem concerns learning and decision making in a game-theoretic context (respectively in §3.4.1 and §3.4.2$_{144}$). Apart from the game-theoretic context, the main new methods we introduce illustrate the use of the theory of imprecise probabilities as a basis for a theory of decision making. The second problem concerns learning the parameters of a Markov chain on the basis of one or more generated state sequences. Markov chains are introduced in §3.4.3$_{146}$ and we propose our inference models for them in §3.4.4$_{149}$.

3.4.1 *Game-theoretic learning*

Quaeghebeur [2001, 2003] and Quaeghebeur & De Cooman [2003*b*, 2009] treat this subject extensively and provide far more context.

Consider the situation where our loyal subject finds himself in a sequential decision problem that can be modeled as a game, with the following characteristics:

(i) It involves two players: himself and his opponent.
(ii) It can consist of an arbitrary number of rounds.
(iii) In any round, his opponent can choose any strategy in a finite set of strategies $\mathscr{X}$; this strategy set is fixed for the whole game.
(iv) With each of his own strategy choices, there corresponds a gamble on $\mathscr{X}$ that represents his possible payoff (winnings or losses).
(v) In any round, both players simultaneously declare the strategy they have chosen for that round, so our subject can then calculate his payoff – measured in some linear precise utility (cf. §1.1.6$_{33}$) – for that round; he can also keep track of his opponent's subsequent strategy choices.

We suppose that our subject wishes to take a structured approach to choosing his own strategies: he wants to model his opponent's behavior and his own uncertainty about that behavior. Based on that model, which he updates whenever he receives new information about his opponent's behavior, he is going to determine in every round which of his strategies are optimal and then play one of those optimal strategies.

The 'fixed mixed strategy'-assumption corresponds to so-called fictitious play in game theory [Brown 1951; Robinson 1951].

To model his opponent's behavior, our subject assumes his opponent is playing a so-called fixed mixed strategy; this means that in every round the opponent uses the same randomization device to choose his strategy. In other words, he assumes his opponent behaves as if he had an urn with colored marbles – where each color corresponds to some

strategy – from which he draws one *with replacement* every round; so he can model his opponent as a multinomial sampling process which is characterized by some unknown frequency vector in $\Delta_{\mathscr{X}}$. He can make this assumption either to simplify matters or because he has some information that supports it. This assumption implies that the sequence of observed strategies is infinitely exchangeable and leads us squarely to the sensitivity analysis interpretation of the theory of imprecise probabilities as a basis of choosing an uncertainty model (cf. §3.3.2$_{140}$): we know there is a precise model, the urn's composition.

Given that the opponent has been conceptually replaced by an urn, what further assumptions do we make about what our subject does or does not know?

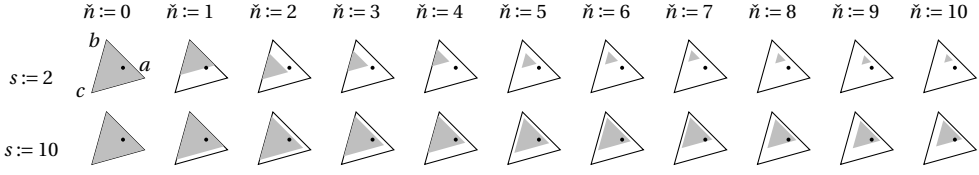
(i) At the start of the game, he has no clue about the setting of the randomization device.

(ii) He evidently assumes that how he pools or labels his opponent's strategies cannot influence the device.

These reasonable-sounding assumptions, together with the infinite exchangeability assumption above and the aptness of the sensitivity analysis interpretation make the IDM an ideal choice for the inference model to help our subject learn from his opponent's behavior.

So our subject chooses some number of pseudocounts s that suits him, dutifully records the sequence of strategies $\check{x}$ his opponent plays through the rounds, calculates the corresponding strategy count vector $\check{m} := C_{\mathscr{X}}\check{x}$, and stands at the ready with his up-to-date inference model $\underline{\mathrm{Di}}(\bullet|\check{m}, s)$ (cf. (3.113)$_{139}$). Ready for what? Ready to determine what strategy to play in the next round, which can be the first round (for which he uses the prior described by (3.114)$_{140}$). How to do this is the subject of the next subsection.

But before we do that, we give a graphical illustration of the evolution of the subject's inferences about his opponent's mixed strategy. This is done by drawing (in gray) the evolution (with $\check{n}$) in the simplex $\Delta_{\mathscr{X}}$ of the set $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$ of parameters that determine the inference model $\underline{\mathrm{Di}}(\bullet|\check{m}, s)$. Take, for this illustration, $\mathscr{X} := \{a, b, c\}$ and consider the observed strategy sequence $\check{x} := bcbbabbaab$ of total length 10.



The fixed mixed strategy I used to generate the strategy sequence was $(\vartheta_a, \vartheta_b, \vartheta_c) := (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$; it is indicated with a black dot. This illustration underlines that one cannot expect the gray set $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$ to contain the black dot mixed strategy. Although increasing the imprecision by in-

creasing s will improve the chance of this happening, the fact that the standard deviation decreases in $\sqrt{\check{n}}$ and the set's volume in $\check{n}^{|\mathscr{X}|-1}$ makes it almost sure not to happen eventually.

3.4.2 *Game theoretic decision making*

We are now going to show how a strategy can be chosen that is in some sense optimal. First of all, we assume that the player has a finite set of strategies $0..i$ at his disposal, where i is a natural number. With each strategy j in $0..i$, there corresponds a payoff function f_j. This is a gamble on the set $\mathscr{X}$ of the opponent's strategies: after choosing a strategy j, the resulting payoff is unknown until the opponent has revealed his strategy choice. The set of all payoff gambles is $\mathscr{K} := \{f_j \mid j : 0..i\}$. Just like his opponent supposedly does, he can use a randomization device – for which he is free to change the setting every round – to choose his strategy for him; i.e., he can play a mixed strategy. With every mixed strategy $\lambda : \Delta_{0..i}$, there corresponds a linear gamble $f_\lambda := \sum_{j:0..i} \lambda_j \cdot f_j$ on $\Delta_{\mathscr{X}}$ that returns the expected payoff under this mixed strategy for every one of his opponent's mixed strategies. We allow our subject to choose any mixed strategy, so – because choosing a strategy implies choosing a payoff gamble – we allow him to use any (expected) payoff gamble in $\text{co}\,\mathscr{K} \subset \mathscr{C}_{\mathscr{X}}$, the convex hull of $\mathscr{K}$.

Illustrating the convex hull operator: $\text{co}\{0,1\} = [0,1]$.

To be able to choose a strategy from $\Delta_{0..i}$, the subject must determine which of his payoff gambles and thus strategies are optimal, so he needs an optimality criterion to guide him [Troffaes 2005, 2007].

The optimality criterion we are going to use first is maximality [Walley 1991, §3.9.2$_{161}$]. This criterion works as follows: Pairwise comparisons are made between all the payoff gambles to determine if one is more optimal than the other, if they are equally optimal, or if their relative optimality is undecidable. On the basis of these pairwise comparisons, a partial ordering of the gambles is made. (Note the relationship with preference orders §1.1.3$_{30}$.) The maximal elements of this order – i.e., those for which no other gamble exists that is more optimal – are *the* optimal payoff gambles, one of which should be chosen for the next round.

Because we are working with a convex set of gambles, maximality is here equivalent to E-admissibility [Levi 1980; Walley 1991, §3.9.6$_{163}$].

We state the pairwise comparison defining maximality first in generic terms: Let Ω be some possibility space, $\mathscr{N} :\subset \mathscr{L}_\Omega$, and $\underline{P}$ some coherent lower prevision on $\mathscr{N}$. We compare two gambles g and h in $\mathscr{N}$ as follows:

(i) g is better than, or dominates, h when $\underline{P}(g-h) > 0$: the supremum acceptable price for swapping h for g must be positive; or, equivalently, using conjugacy (1.14)$_{36}$,
(ii) g is worse than, or is dominated by, h when $\overline{P}(g-h) < 0$;
(iii) g and h are equally good when $\underline{P}(g-h) = 0 = \overline{P}(h-g)$: one stays indifferent to the swapping of g for h;
(iv) g and h are incomparable when none of the above holds: one does not have enough information to decide between the two gambles.

According to the maximality criterion, those gambles in $\mathscr{N}$ that are undominated are optimal. This means that a gamble g is optimal in $\mathscr{N}$ when $\min_{h:\mathscr{N}} \overline{P}(g-h) \geq 0$, where the inequality can of course immediately be replaced by an equality (take $h := g$).

Returning to our game-theoretic context, our subject has a parametric uncertainty model $\underline{\mathrm{Di}}(\bullet|\check{m}, s)$ for all continuous gambles on the possibility space $\Delta_{\mathscr{X}}$ of all the possible mixed strategies his opponent can use. Our subject has to find the optimal payoff gambles *for the coming round*, so he has to use the immediate prediction model corresponding to $\underline{\mathrm{Di}}(\bullet|\check{m}, s)$. This is (cf. $(3.115)_{141}$; let f be a gamble on $\mathscr{X}$)

$$\overline{\mathrm{Di}}\big(\mathrm{Mn}(f|1,\bullet) \,\big|\, \check{m}, s\big) = \overline{\mathrm{Dm}}\big(\mathrm{Mh}(f|\bullet) \,\big|\, 1, \check{m}, s\big) = \overline{\mathrm{Wa}}(f|\check{m}, s).$$

Under the maximality criterion, a payoff gamble g is optimal in $\mathrm{co}\,\mathscr{K}$ when $\min_{h:\mathrm{co}\,\mathscr{K}} \overline{\mathrm{Wa}}(g-h|\check{m}, s) = 0$, i.e., nonnegative. We can rewrite the left-hand side, by first writing $\overline{\mathrm{Wa}}(\bullet|\check{m}, s)$ as a lower envelope $(3.102)_{135}$ and then using the minimax theorem [Walley 1991, §E6_{613}], which can be applied without worries as both $\mathrm{co}\,\mathscr{K}$ and $\Delta_{\mathscr{X}}$ are compact, convex sets and $\mathrm{Wa}(\bullet|\bullet)$ is linear in both arguments:

$$\begin{aligned}
&\min_{h:\mathrm{co}\,\mathscr{K}} \overline{\mathrm{Wa}}(g-h|\check{m}, s) \\
&\qquad = \min_{h:\mathrm{co}\,\mathscr{K}} \max_{t:\Delta_{\mathscr{X}}} \mathrm{Wa}(g-h|\check{m}+s\cdot t) \\
&\qquad = \max_{t:\Delta_{\mathscr{X}}} \min_{g:\mathrm{co}\,\mathscr{K}} \mathrm{Wa}(g-h|\check{m}+s\cdot t) \\
&\qquad = \max_{t:\Delta_{\mathscr{X}}} \big(\mathrm{Wa}(g|\check{m}+s\cdot t) - \max_{h:\mathrm{co}\,\mathscr{K}} \mathrm{Wa}(h|\check{m}+s\cdot t)\big).
\end{aligned}$$

We see that this expression will be zero and thus g will be optimal, whenever there is a frequency vector t for which g maximizes the expected utility under $\mathrm{Wa}(\bullet|\check{m}+s\cdot t)$. Every frequency vector determines an expected mixed strategy $\frac{\check{m}+s\cdot t}{\check{n}+s}$ of the opponent, so we can write the set of mixed strategies the subject expects to be possible choices for his opponent as $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$. Incorporating this piece of interpretation, we can formally write down the set of optimal strategies as

$$\bigcup_{t:\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}} \arg\max_{\lambda:\Delta_{0..j}} \mathrm{Wa}(f_\lambda|t). \tag{3.116}$$

Prepending arg in front of an extremum operator (e.g., min, max) returns the set of arguments that extremizes the expression. For example, $\mathbb{Z} = \arg\min\{|\sin(\pi\cdot\bullet)|\}$.

In game-theoretic terminology [see, e.g., Friedman 1989] this would be called a best reply to $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$.

The maximality optimality criterion only takes into account the uncertainty model of our subject and the set of strategies (and thus gambles) he can choose from. His attitude towards the game and his opponent or information (he thinks) he has about the game and his opponent that was not included in his uncertainty model – e.g., because it was irrelevant for that purpose or just too complicated to do – can influence the choice of optimality criterion. As a further step, we consider the case that our subject thinks his opponent chooses his strategy to minimize

our subject's payoff; he may think this because he is a pessimist, or because he thinks his opponent is evil, or because he knows that the game he is playing is strictly competitive (each round, the sum of both players' payoff is constant; zero-sum games are the archetypical example). In such a context, it is reasonable for our subject to try to minimize his losses. He can do this by choosing to play a strategy with a gamble that has the best expected worst-case payoff: he will try to maximize his lower prevision.

We can formalize this by saying that a payoff gamble g is optimal when $\underline{\mathrm{Wa}}(g|\check{m},s) = \max_{h:\mathrm{co}\mathscr{K}} \underline{\mathrm{Wa}}(h|\check{m},s)$. Using similar basic manipulations as in the case of maximality, the set of optimal strategies becomes

$$\arg\max_{\lambda:\Delta_{0..j}} \min_{t:\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}} \mathrm{Wa}(f_\lambda|t), \tag{3.117}$$

which can be seen to be a subset of the set of strategies (3.116)$_\frown$ optimal under maximality. These optimal strategies are the so-called maximin strategies for $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$: their gambles maximize the expected payoff, given that the opponent is expected to choose a strategy in $\frac{\check{m}+s\cdot\Delta_{\mathscr{X}}}{\check{n}+s}$ that makes this expected payoff as small as possible.

This concludes our first application of the IDM.

3.4.3 *Markov chains: linking multinomial processes for fun and profit*

Kemeny & Snell [1976] give a good introduction to finite Markov chains.

The second application of the IDM we are going to look at is learning (stationary) finite Markov chains, a widely used type of stochastic process [see, e.g., Dayhoff et al. 1978; Sarukkai 2000]. This means that we wish to learn (about) the parameters that completely describe this stochastic process on the basis of its output. In this subsection, we introduce Markov chains, their parameters, and their likelihood functions; in the next subsection we propose two related inference models.

A stationary finite Markov chain can be seen as a discrete stochastic process, i.e. as an infinite length sequence X of random variables. Each of the random variables can take a value within some finite set $\mathscr{X}$ of states (or categories). The central assumption that makes a discrete stochastic process a Markov chain is that the chance that $X_k = z$ – where $k:\mathbb{N}_{>1}$ and $z:\mathscr{X}$ – is *only* influenced by the state (that was observed) for X_{k-1} and not on the realization of any 'previous' random variable, i.e., with indices smaller than $k-1$; this is called the Markov condition or Markov property. In more formal language, we can say that conditional on the value that X_{k-1} assumes, the random variable X_k is independent from the random variables X_j, with j in $\mathbb{N}_{<k-1}$. (We call the Markov model stationary to stress that there is no dependence on the sequence index k.)

The Markov condition makes it natural to think in terms of transitions between states, which can be represented by pairs of states. Although imprecise Markov chains can be defined and used [De Cooman

et al. 2008*a*, 2009*a*; Dhaenens 2007; Škulj 2007], we assume here that the observations are generated by a precise Markov chain. So, as in the previous subsection, we place ourselves squarely in a sensitivity analysis context.

A precise Markov chain is usually parameterized using a (stochastic) transition matrix $\Theta : (\Delta_{\mathscr{X}})^{\mathscr{X}}$ that contains the transition probabilities: for every state z, the row θ_z^{T} contains the probabilities $\theta_{zz'}$ for a transition from state z to the next state $z' : \mathscr{X}$. So our aim is to build an inference model that tells us something about these transition probabilities based on the observed transitions.

Matrix and vector transposition is denoted by $\bullet^{\mathsf{T}}$.

When we want to be able to deal with multiple state sequences, we also need a model for how the initial state of these sequences is chosen. We assume the initial states of the different sequences to form an exchangeable sequence generated by an unknown multinomial distribution. This so-called initial distribution is parameterized by a frequency vector ϑ in $\Delta_{\mathscr{X}}$ that describes the composition of the proverbial urn from which the initial states are drawn with replacement.

When we observe a partial state sequence $\check{x} : \mathscr{X}^{\check{n}}$ of length $\check{n}+1$ and indexed from 0 to $\check{n}$, where $\check{n} : \mathbb{N}_{>0}$ is the number of observed transitions, we can associate with it a count matrix $\check{M} : \check{n}^{\mathscr{X}\times\mathscr{X}}$. For this, we first define, for every state z and every sequence $\check{x}$, the subsequence

$$\check{x}_z := (\check{x}_{k+1} | k : 0..\check{n}-1 \wedge \check{x}_k = z) \tag{3.118}$$

consisting of those states that immediately follow state z in the sequence $\check{x}$. Then the row of $\check{M}$ consisting of transitions out of state z is $\check{m}_z^{\mathsf{T}} := C_{\mathscr{X}} \check{x}_z$ and the number of observed transitions from z to z' is $\check{m}_{zz'}$.

Because it will be of use later on, we also introduce the frequency matrix $\check{F} : \big(\Delta_{\mathscr{X}} \cup (\mathscr{X} ; 0)\big)^{\mathscr{X}}$, for which each row $\check{f}_z^{\mathsf{T}}$ is defined by $\check{f}_z = \frac{\check{m}_z}{\sum \check{m}_z}$ when $\sum \check{m}_z > 0$ and is identically zero otherwise. This matrix $\check{F}$ is stochastic in the sense that each nonzero row's components are nonnegative and sum to one.

Due to the Markov condition, the sequence $\check{x}_z$ with count vector $\check{m}_z$ is a multinomial sample. The corresponding sequence and count likelihood functions are $B_{\check{x}_z}$ and $B_{\check{m}_z}$, respectively (cf. $(3.34)_{107}$ and $(3.25)_{104}$).

The likelihood function corresponding to a Markov chain is composed of the multinomial likelihoods for each of the different states *and* the likelihood of observing the initial state $\check{x}_0$. Let $\check{x}$ be a state sequence with count and frequency matrices $\check{M}$ and $\check{F}$ generated by a Markov chain with transition matrix Θ and initial distribution ϑ, then the probability of encountering this state sequence is

$$W_{\check{x}}(\vartheta, \Theta) := \vartheta_{\check{x}_0} \cdot \prod_{k:1..\check{n}} \theta_{\check{x}_{k-1}\check{x}_k} = \vartheta_{\check{x}_0} \cdot \prod_{z:\mathscr{X}} B_{\check{x}_z}\theta_z, \tag{3.119}$$

which gives rise to (what we call) a Markov linear prevision – or rather,

one for every possible number of transitions $n : \mathbb{N}_{>0}$ – defined for any gamble f on $\mathscr{X}^n$ by

$$\mathrm{Mv}(f|n,\bullet,\bullet) := \sum_{y:\mathscr{X}^{n+1}} f\,y \cdot W_y. \tag{3.120}$$

This prevision is in some sense composed of multinomial previsions (one for each state and one for the initial distribution). Due to the fact that the length of the subsequences of states transitioning out of a specific state – the length parameter for the multinomial for that state – varies for different state sequences, writing it down in terms of multinomial previsions is hard.

We have learned so far that $W_{\check{x}}$ is the Markov likelihood function corresponding to the state sequence $\check{x}$. Because $B_{\check{m}_z} \propto B_{\check{x}_z}$, this likelihood is the same for every state sequence with the same count matrix starting in the same initial state. This means that the pair $(\check{x}_0, \check{M})$ is a sufficient statistic and the order information available in the state sequence $\check{x}$ not expressed by this pair is ancillary, i.e., irrelevant for updating according to the discrete likelihood principle (cf. §3.1.6$_{115}$).

Although we know that the likelihood function corresponding to the observation of some initial state $\check{x}_0$ and some transition count matrix $\check{M}$ will be proportional to $W_{\check{x}}$, it is interesting to know this proportionality constant; i.e., to know how many different state sequences correspond to a given transition count matrix generated by a Markov chain starting in a given initial state. We know that one factor in this will be the product of all the proportionality constants between $B_{\check{m}_z}$ and $B_{\check{x}_z}$ for all the different states z, which express how many conditional state sequences correspond to each of the matrix's rows. The fact that in a Markov chain, after transitioning into a state, one can by definition only transition out of that same state, means we need another factor to express that a valid state sequence is not just any collection of valid conditional state sequences. Whittle [1955] found that this factor is $|\mathbb{1} - \check{F}|_{\check{x}_{\check{n}}\check{x}_0}$, the $(\check{x}_{\check{n}}, \check{x}_0)$-cofactor of the matrix $\mathbb{1} - \check{F}$. In this expression, the final state $\check{x}_{\check{n}}$ appears; it can be derived from $\check{x}_0$ and $\check{M}$, as all states are transitioned into as many times as they are transitioned out of, except possibly the initial and final states: for all states z, it holds that $\sum \check{m}_z - \sum \check{m}_{\bullet z} = \delta_{\check{x}_0 z} - \delta_{z\check{x}_n}$.

The identity matrix is denoted by $\mathbb{1}$.

To see that a compensating factor such as $|\mathbb{1} - \check{F}|_{\check{x}_n\check{x}_0}$ is necessary, consider the following example: Take $\mathscr{X} := \{a, b\}$, $\check{M} := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, and $\check{x}_0 := b$, which implies that $\check{F} = \frac{1}{2}\check{M}$ and $\check{x}_n = b$. Seen as two sets of multinomial transition samples, there are four ways to order them: aa, ab or ab, aa and ba, bb or bb, ba. The (b, b)-cofactor of $\mathbb{1} - \frac{1}{2}\check{M}$ is $(\mathbb{1} - \frac{1}{2}\check{M})_{aa} = \frac{1}{2}$, so there are only two possible state sequences, i.e., only two of the four multinomial transitions samples can be patched together: ba, aa with ab, bb to $baabb$ and ba, aa with bb, ab to $bbaab$.

Combining the proportionality constants discussed in the next-to-last paragraph with the expression for the likelihood function (3.119)$_\frown$ of a

state sequence, we can write down the likelihood function corresponding to a given count matrix $\check{M}$ and initial state $\check{x}_0$ [also see Billingsley 1961]:

$$W_{\check{x}_0,\check{M}}(\vartheta,\Theta) := \vartheta_{\check{x}_0} \cdot |\mathbb{1} - \check{F}|_{\check{x}_{\hat{n}}\check{x}_0} \cdot \prod_{z:\mathscr{X}} B_{\check{m}_z}\theta_z. \tag{3.121}$$

The most interesting thing we have now learned is that not all matrices in $n^{\mathscr{X}\times\mathscr{X}}$ – where $n:\mathbb{N}_{>0}$ – are transition count matrices resulting from a state sequence with n transitions. We formalize the Markovian constraint on these count matrices, which guarantees that they result from a state sequence starting in z, with a predicate $z\text{-}\mathrm{mkv}:n^{\mathscr{X}\times\mathscr{X}} \to \mathbb{B}$ defined for every such matrix $M:n^{\mathscr{X}\times\mathscr{X}}$ by

$$\begin{aligned} z\text{-}\mathrm{mkv}\, M \Leftrightarrow & \textstyle\sum m_z > 0 \\ & \wedge \exists z':\mathscr{X}; \forall z'':\mathscr{X}; \textstyle\sum \check{m}_{z''} - \sum \check{m}_{\bullet z''} = \delta_{zz''} - \delta_{z''z'}. \end{aligned} \tag{3.122}$$

Use this to define the state-agnostic predicate $\mathrm{mkv}:n^{\mathscr{X}\times\mathscr{X}} \to \mathbb{B}$ with

$$\mathrm{mkv}\, M \Leftrightarrow \exists z:\mathscr{X}; z\text{-}\mathrm{mkv}\, M. \tag{3.123}$$

This state-agnostic predicate allows us to give expressions for the Markov linear previsions for count matrices (one for every total transition count $n:\mathbb{N}_{>0}$). It is defined for any gamble h on $(n^{\mathscr{X}\times\mathscr{X}})_{\mathrm{mkv}}$ by

$$\mathrm{Cv}(h|n,\bullet,\bullet) := \textstyle\sum_{z:\mathscr{X}} \sum_{M:(n^{\mathscr{X}\times\mathscr{X}})_{z\text{-}\mathrm{mkv}}} hM \cdot W_{z,M}. \tag{3.124}$$

Similarly to the sequence Markov prevision (3.120), this prevision is composed of count-multinomial previsions (one for each state and one for the initial distribution). Again, it is hard to write it down as such.

3.4.4 *Learning Markov chains*

Quaeghebeur & De Cooman [2003*a*] sketches how an imprecise-probabilistic inference model for finite Markov chains could be used for command-line completion.

Now that we have familiarized ourselves with the parameters, likelihood functions, and previsions of Markov chains, we are ready to propose inference models for this type of stochastic process.

Because we already know that the process satisfies the Markov condition, there is no uncertainty about how the sequence of random variables X can be split up into infinitely exchangeable subsequences X_z of transitions out of the different states z. The only thing there is uncertainty about is the relative likelihoods of these transitions. So, assuming representation insensitivity, i.e., that

(i) we start from a state of prior ignorance about these relative likelihoods,
(ii) the states' names are irrelevant, and
(iii) the grouping of states does not modify the relative likelihoods,

and considering we are working in a sensitivity analysis context, it is reasonable to use one IDM per state as the inference model for the unknown probabilities for transitions out of that state. For example, the IDM for transitions out of state z after observing the count matrix $\check{M}$ – or after

a string of observations for which the count matrices sum up to $\check{M}$ – is $\underline{\mathrm{Di}}(\bullet|\check{m}_z, s_z)$, where we wish to draw attention to the fact that we can choose the number of pseudocounts $s_z : \mathbb{R}_{>0}$ to be z-dependent.

Concerning the initial distribution, we have seen in the previous subsection that it is assumed to be an unknown multinomial distribution, for which we will use an IDM as well. Let $\check{\mathfrak{m}}$ be the count vector corresponding to the finite sequence of observed initial states and let $\mathfrak{s}$ be the number of pseudocounts chosen for learning the initial state, then the IDM used is $\underline{\mathrm{Di}}(\bullet|\check{\mathfrak{m}}, \mathfrak{s})$.

Of course, the individual state-dependent IDMs and the IDM for the initial state need to be combined into a global inference model for the whole Markov chain. Given our sensitivity analysis context [Walley 1991, §9.1.5$_{446}$] and the assumed independence of these models resulting from the Markov condition, we use a type-1 product (1.88)$_{64}$. All the marginals – the individual state-dependent IDMs and the IDM for the initial state – are lower envelopes of Dirichlet previsions. To calculate the type-1 product, we first need to take products of these Dirichlet linear previsions, so let T be some 'prior' transition matrix in $(\Delta_{\mathscr{X}})^{\mathscr{X}}$ with rows $t^{\intercal}$, s a vector of pseudocounts in $(\mathbb{R}_{>0})^{\mathscr{X}}$, $\mathfrak{t}$ a 'prior' frequency vector in $\Delta_{\mathscr{X}}$ for the initial distribution, then the corresponding precise product prevision, which we call a precise Markov chain Dirichlet model (or PMCDM), is

$$\mathrm{MDi}\big(\bullet \,\big|\, \check{\mathfrak{m}} + \mathfrak{s}\cdot\mathfrak{t}, \check{M} + s\cdot T\big) := \mathrm{Di}(\bullet|\check{\mathfrak{m}} + \mathfrak{s}\cdot\mathfrak{t}) \times \mathrm{Di}^{\times}\big(\bullet \,\big|\, \check{M} + s\cdot T\big), \quad (3.125)$$

where

$$\mathrm{Di}^{\times}\big(\bullet \,\big|\, \check{M} + s\cdot T\big) := \bigtimes_{z:\mathscr{X}} \mathrm{Di}(\bullet|\check{m}_z + s_z\cdot t_z) \quad (3.126)$$

and where $s\cdot T$ is a row-wise product, i.e., the z-row $(s\cdot T)_z$ of the product is $s_z\cdot t_z^{\intercal}$. The precise Markov chain Dirichlet model is defined for all continuous gambles on $\Delta_{\mathscr{X}} \times (\Delta_{\mathscr{X}})^{\mathscr{X}}$. The imprecise Markov chain Dirichlet model (or IMCDM) is then defined as a lower envelope of precise Markov chain Dirichlet models:

$$\underline{\mathrm{MDi}}\big(\bullet \,\big|\, \check{\mathfrak{m}}, \mathfrak{s}, \check{M}, s\big) := \min_{\mathfrak{t}:\Delta_{\mathscr{X}}} \min_{T:(\Delta_{\mathscr{X}})^{\mathscr{X}}} \mathrm{MDi}\big(\bullet \,\big|\, \check{\mathfrak{m}} + \mathfrak{s}\cdot\mathfrak{t}, \check{M} + s\cdot T\big). \quad (3.127)$$

In the IMCDM's current definition, we have to separately choose a pseudocounts value s_z for each state z. We can simplify this by letting the pseudocounts be equal for all states (the notation of definition (3.127) can be kept, but now with s as a positive real number). However, the choice of state space can be somewhat arbitrary itself, which means that for different initial choices for the state space, we can either keep the individual number of pseudocounts per state constant or the total number for all the states, but not both.

The former option, using a IMCDM with pseudocounts that are con-

stant over the states, currently seems like an entirely reasonable choice to me: The pseudocounts parameter arose in our derivation of the IDMM (cf. §3.2.5$_{131}$), where we saw that it characterized the mixing sequences, i.e., it characterized how fast we moved from a vacuous system to a Haldane system for immediate prediction. Given the current state, immediate prediction for a Markov chain is exactly the same as immediate categorical prediction, so there is no clear reason for me to let the pseudocounts depend on the number of other states – which has no relevance for the immediate prediction.

However, due to where it appears in the formulas defining the ID(M)M (cf. (3.105)$_{135}$ and (3.113)$_{139}$), the pseudocounts parameter is often given the intuitive interpretation of a hypothetical sample. Letting the total number of pseudocounts for the IMCDM depend on the somewhat arbitrary choice of the total number of states may seem contradictory in the context of this intuitive interpretation. The question then becomes: is it possible to modify the IMCDM in such a way that we obtain an inference model for learning Markov chains that is compatible with this interpretation? The answer is yes.

When keeping the total number of pseudocounts constant, we need to distribute this total number among the different states. Of course, in a state of prior ignorance, we have no clue as to how the hypothetical sample was obtained and thus how it should be distributed over the different states. Luckily for us, users of the theory of imprecise probabilities, we have the vacuous lower prevision to model this state of knowledge. So, now letting $s : \mathbb{R}_{>0}$ denote the constant total number of pseudocounts over all states, the corresponding parametric inference model becomes

$$\underline{\mathrm{MDj}}\big(\bullet \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s\big) := \inf_{r:\Delta_{\mathscr{X}}} \underline{\mathrm{MDi}}\big(\bullet \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s \cdot r\big). \tag{3.128}$$

An infimum is used because the IMCDM is not defined when for some state both $\check{m}_z = 0$ and $r_z = 0$, which can happen for r on the border of the simplex.

Both the constant per-state pseudocount model $\underline{\mathrm{MDi}}\big(\bullet \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s\big)$ defined by (3.127) and $\underline{\mathrm{MDj}}\big(\bullet \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s\big)$, the constant total pseudocount model defined by (3.128) derived from it, can be used to define predictive sequence and count inference models.

Assume first that we are interested in predictions about a future state sequence with $\hat{n} : \mathbb{N}_{>0}$ transitions, i.e., of length $\hat{n}+1$, then the predictive inference models will be

$$\underline{\mathrm{MDi}}\big(\mathrm{Mv}(\bullet|\hat{n}, \bullet, \bullet) \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s\big) \quad \text{and} \quad \underline{\mathrm{MDj}}\big(\mathrm{Mv}(\bullet|\hat{n}, \bullet, \bullet) \,\big|\, \check{m}, \mathfrak{s}, \check{M}, s\big).$$

Both models are defined for all gambles on $\mathscr{X}^{\hat{n}+1}$ (cf. (3.120)$_{148}$). When we are interested in predictions about a future $\hat{n}$-transition count matrix,

they will be

$$\underline{\mathrm{MDi}}\bigl(\mathrm{Cv}(\bullet|\hat{n},\bullet,\bullet)\,\big|\,\check{\mathfrak{m}},\mathfrak{s},\check{M},s\bigr) \quad \text{and} \quad \underline{\mathrm{MDj}}\bigl(\mathrm{Cv}(\bullet|\hat{n},\bullet,\bullet)\,\big|\,\check{\mathfrak{m}},\mathfrak{s},\check{M},s\bigr),$$

both defined for all gambles on $(\hat{n}^{\mathscr{X}\times\mathscr{X}})_{\mathrm{mkv}}$ (cf. $(3.124)_{149}$).

The immediate prediction models

$$\underline{\mathrm{MDi}}\bigl(\mathrm{Mv}(\bullet|1,\bullet,\bullet)\,\big|\,\check{\mathfrak{m}},\mathfrak{s},\check{M},s\bigr) \quad \text{and} \quad \underline{\mathrm{MDj}}\bigl(\mathrm{Mv}(\bullet|1,\bullet,\bullet)\,\big|\,\check{\mathfrak{m}},\mathfrak{s},\check{M},s\bigr)$$

can be used to learn imprecise Markov chains [see, e.g., De Cooman et al. 2008*a*, 2009*a*]: they provide predictions about an unobserved future initial state and one ensuing transition. The marginal for the initial state – which is identical for both prediction models – is $\underline{\mathrm{Wa}}(\bullet|\check{\mathfrak{m}},\mathfrak{s})$; the updated lower prevision for a given initial state z is $\underline{\mathrm{Wa}}(\bullet|\check{m}_z,s_z)$ or $\underline{\mathrm{Wa}}(\bullet|\check{m}_z,s)$. These updated lower previsions are obtained by regular extension $(1.70)_{58}$. Let $\mathfrak{t}$ be some frequency vector in $\Delta_{\mathscr{X}}$, T some Markov matrix in $(\Delta_{\mathscr{X}})^{\mathscr{X}}$, $\hat{x}_0$ some initial state in $\mathscr{X}$, and f a gamble on $\mathscr{X}$; to calculate these regular extensions, we used the fact that (cf. $(3.125)_{150}$)

$$\mathrm{MDi}\Bigl(\mathrm{Mv}\bigl(I^{\iota\hat{x}_0\times\mathscr{X}}\cdot f\,\big|\,1,\bullet,\bullet\bigr)\,\Big|\,\check{\mathfrak{m}}+\mathfrak{s}\cdot\mathfrak{t},\check{M}+s\cdot T\Bigr) = \frac{\check{\mathfrak{m}}_{\hat{x}_0}+\mathfrak{s}\cdot\mathfrak{t}_{\hat{x}_0}}{\sum\check{\mathfrak{m}}+\mathfrak{s}}\cdot\mathrm{Wa}(f|\check{m}_{\hat{x}_0}+s\cdot t_{\hat{x}_0}),$$

because (let $\vartheta:\Delta_{\mathscr{X}}$ and $\Theta:(\Delta_{\mathscr{X}})^{\mathscr{X}}$; cf. $(3.120)_{148}$)

$$\mathrm{Mv}\bigl(I^{\iota\hat{x}_0\times\mathscr{X}}\cdot f\,\big|\,1,\vartheta,\Theta\bigr)=\vartheta_{\hat{x}_0}\cdot\mathrm{Wa}(f|\theta_{\hat{x}_0}).$$

We use regular extension here because when $\check{\mathfrak{m}}_{\hat{x}_0}=0$, we do not want the updated lower prevision to automatically become vacuous.

The updated previsions also provide us with conservative interval estimates $[\frac{\check{m}_{zz'}}{\sum\check{m}_z+s},\frac{\check{m}_{zz'}+s}{\sum\check{m}_z+s}]$, for every component $(z,z':\mathscr{X}^2)$ of the transition matrix of the Markov chain we are learning about. If the Markov chain is regular, this forms a consistent sequence of estimators for increasing $\check{n}$ [Quaeghebeur 2004], i.e., it converges in probability to the true transition matrix.

With this, we end this chapter that has dealt exclusively with finite sample spaces. In the next chapter we consider infinite sample spaces as well.


SOIS
BEZ

INFERENCE MODELS FOR EXPONENTIAL FAMILIES

> "Ah," disse il professor Camestres, "come già si diceva nel primo *Liber legis*, ogni numero è infinito, e non c'è differenza!"
>
> "Capisco," disse Belbo. "Ma non pensa che tutto questo sia un poco oscuro per il lettore comune?"
>
> Camestres quasi sobbalzò sulla sedia. "Ma è assolutamente indispensabile. Chi comprendesse questi segreti senza la dovuta preparazione precipiterebbe nell'Abisso! Già nel renderli pubblici in modo velato io corro dei rischi, mi credano. Io mi muovo nell'ambito dell'adorazione della Bestia, ma in modo più radicale di Crowley, [...]"
>
> Eco [1988, Ch. 44]

We wish to learn from samples: make predictions about future samples or draw conclusions about the process generating the samples. This is respectively called predictive and parametric inference. The uncertainty model making these inferences based on the given samples is called an inference model. This was our stated goal at the beginning of the previous chapter, 'Inference models'$_{94}$, and it still is.

Whereas in the previous chapter we focused on inference for categorical sampling, we here expand our view to learning from samples that come from infinite discrete or continuous spaces. Whereas in the previous chapter we carefully built up the inference models step-by-step, each time explicitly mentioning the assumptions used, we here propose – analogously to how the IDM was first proposed [Walley 1996] – a class of imprecise-probabilistic inference models as generalizations of a well-known class of Bayesian inference models.

Quaeghebeur & De Cooman [2005] compactly present imprecise-probabilistic inference models for exponential families.

The Bayesian inference models we are going to generalize are the conjugate inference models [Bernardo & Smith 1994, §5.2$_{265}$]; but conjugate to what? Recall from the subsection §3.3.2$_{140}$ on conjugate updating that there is a class of sampling models described by likelihood functions that have some very interesting properties in common with the multinomial likelihood that was central in the previous chapter. This is the class of regular exponential likelihood functions; it will be central to this chapter.

Raiffa & Schlaifer [1968] first introduced conjugate priors; Fink [1995] gives a compendium.

The first thing we will do, in §4.1, is extensively familiarize ourselves with the regular exponential families of likelihoods and their classical Bayesian conjugate parametric and predictive inference models. Once this is done and digested with the help of some examples, you are ready

to follow in my footsteps and take the by now perhaps straightforward (but not necessarily easy) generalization step of §4.2$_{173}$, where we go from inference models made up of single linear previsions to imprecise-probabilistic ones consisting of sets of linear previsions. The last part of the chapter, §4.3$_{180}$, is meant to show how these specific imprecise-probabilistic inference models can be applied using the example of naive credal classification. To supplement the specific exponential families and their derived inference models given as examples in this chapter, we have added the Bestiarium$_{202}$, an appendix containing a similar treatment for a number of additional exponential families.

4.1 Exponential families & friends

In this section, we gather and review information about exponential families and related probabilistic objects that can be found in the literature. It serves as a basis for the next section, where we propose imprecise-probabilistic parametric and predictive inference models for exponential families. The theoretical subsections §4.1.1, §4.1.4$_{164}$ and §4.1.5$_{166}$ are complemented by examples in §4.1.2$_{159}$, §4.1.3$_{160}$, §4.1.6$_{168}$, and §4.1.7$_{171}$.

When talking about exponential families, we use the terminology of Barndorff-Nielsen [1978, Ch. 8$_{111}$]. Brown's [1986] monograph is another interesting source.

4.1.1 *Regular exponential families*

Consider a subject who is making an arbitrary number of observations of a certain phenomenon. These samples can take a value in some sample space we generically denote by $\mathscr{X}$. In this chapter, contrary to the previous one, we consider both finite and countably or uncountably infinite sample spaces. So formally, what we consider is an infinite sequence $X := (X_k \mid k : \mathbb{N}_{>0})$ of random variables drawn from the same sample space $\mathscr{X}$. (These random variables can be scalar or vectorial.)

We assume that the draws are identically distributed according to some unknown member of a known regular exponential family and that the draws are independent conditional on the knowledge of the actual distribution. For example, we could assume that the draws come from a fixed normal distribution with unknown mean and variance. This assumption holds – under an additional σ-additivity assumption – when the sequence of random variables is assumed to be infinitely exchangeable [Kallenberg 2005, Thm 1.1$_{177}$], which is often considered reasonable.

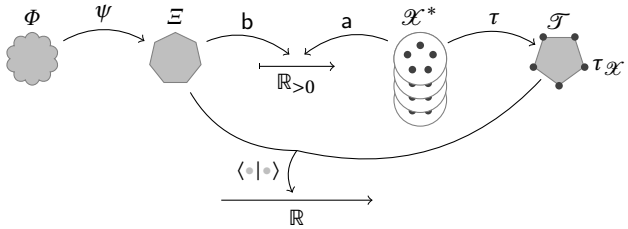
Why do we restrict ourselves to (or consider as broad a class as) the regular exponential families? The main reason is that these families and their conjugate families have so much in common with, respectively, the multinomial distributions (which also form an exponential family) and the Dirichlet distribution, that proposing inference models similar to the IDM is actually quite straightforward. One important property of exponential family likelihoods is that they are parameterized by sufficient statistics of finite dimension.

Among those that satisfy some regularity conditions, exponential family likelihoods are actually the only ones having finite dimensional sufficient statistics [DeGroot 2004, §9.2$_{161}$].

For every observed sample $z : \mathscr{X}$ generated by some exponential family distribution, we can write out the corresponding likelihood function $E_z^{\mathsf{a},\tau,\psi} : \Phi \to \mathbb{R}_{>0}$. It is fully characterized – including the domain and the set of parameters Φ – by the three functions, $\mathsf{a} : \mathscr{X}^* \to \mathbb{R}_{>0}$, $\tau : \mathscr{X}^* \to \mathscr{T}$, and $\psi : \Phi \to \Xi$, whose role and ranges will be clarified below. What is the general expression of such an exponential family likelihood function? Let ϕ be some parameter in Φ, then

$$E_z^{\mathsf{a},\tau,\psi}\phi = \mathsf{b}(\psi\phi) \cdot \exp\langle \psi\phi | \tau z \rangle \cdot \mathsf{a}z, \tag{4.1}$$

where $\mathsf{b} : \Xi \to \mathbb{R}_{>0}$ is a normalization function and $\langle \bullet | \bullet \rangle : \Xi \times \mathscr{T} \to \mathbb{R}$ denotes a scalar product between the set of so-called canonical parameters Ξ and the so-called set of means $\mathscr{T}$. The middle factor explains the epithet 'exponential family'. A diagram containing the functions and sets involved in the above definition can help us form a mental picture:



The choice of shapes (convex or not) for the sets and the relationship between $\mathscr{X}^*$ and $\mathscr{T}$ will become clear later on in this subsection.

Gambles on discrete spaces are always measurable.

With likelihoods of the type (4.1), there correspond linear previsions $\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\bullet|\phi)$ on the set $\check{\mathscr{L}}_{\mathscr{X}}$ of all *measurable* gambles. Let f be a gamble in $\check{\mathscr{L}}_{\mathscr{X}}$; for discrete $\mathscr{X}$ (using the counting measure) this prevision is defined by

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(f|\phi) = \textstyle\sum_{z:\mathscr{X}} f z \cdot E_z^{\mathsf{a},\tau,\psi}\phi \tag{4.2}$$

and for continuous $\mathscr{X}$ (using Lebesgue measure) by

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(f|\phi) = \textstyle\int_{\mathscr{X}} f z \cdot E_z^{\mathsf{a},\tau,\psi}\phi \,\mathrm{d}z. \tag{4.3}$$

A set $\left\{\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\bullet|\phi) \,\middle|\, \phi : \Phi\right\}$ is what we call an (indexed) exponential family (of linear previsions). The family is fully characterized by a, τ, and ψ; given a family, each family member is fully characterized by its parameter ϕ.

Let us now discuss the functions appearing in (4.1):

Burrill [1972] is a good reference for measure-theoretic concepts.

(i) In the literature [see, e.g. Barndorff-Nielsen 1978, Ch. 8_{111}], the basis of any exponential family is a σ-finite measure (so exponential family distributions are σ-additive). We only consider Lebesgue or counting measure modified by a (measurable) density or mass function a, which is assumed to be strictly positive, because oth-

erwise we could simplify the model by removing those possible observations that are considered to be practically impossible.

(ii) As is typical in the literature, we take $\mathcal{T}$ and Ξ to be subsets of the same finite-dimensional vector space $\mathbb{R}^d$ on which we place the Euclidean topology; d is a nonzero natural number. This means that both τ and ψ are d-dimensional vector functions.

(iii) We assume that the exponential family is minimal, in the sense that the dimension d is the smallest one that allows $\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\bullet|\phi)$ to be written as it is. A necessary and sufficient condition for this is that the components of τ (which must of course be measurable) and ψ are affinely independent [Barndorff-Nielsen 1978, Corrolary 8.1_{113}].

(iv) We see that $E_z^{\mathsf{a},\tau,\psi}\phi$ only depends on ϕ through ψ: the parameter space Φ is only used for interpretational purposes, the mathematically relevant space is the space Ξ of canonical parameters. Whenever $\psi = \mathrm{id}_\Phi$ and thus $\Phi = \Xi$, we are dealing with a so-called canonical exponential family. For these, we can drop the superscript ψ; so $E_z^{\mathsf{a},\tau}$ is the canonical family likelihood function, related to a corresponding noncanonical one by

$$E_z^{\mathsf{a},\tau,\psi}\phi = E_z^{\mathsf{a},\tau}(\psi\phi).$$

We often discuss the properties of an exponential family in terms of the corresponding canonical one.

(v) The normalization function b is defined for every canonical parameter ξ in Ξ by $\mathrm{Ef}^{\mathsf{a},\tau}(\mathscr{X}|\xi) = 1$, i.e.,

$$\mathsf{b}\xi = 1/\textstyle\sum_{z:\mathscr{X}} \exp\langle\xi|\tau z\rangle\cdot\mathsf{a}z \quad \text{or} \quad \mathsf{b}\xi = 1/\int_{\mathscr{X}} \exp\langle\xi|\tau z\rangle\cdot\mathsf{a}z\,\mathrm{d}z.$$

(vi) The set of canonical parameters Ξ is chosen such that it contains all mathematically acceptable parameters. To wit, it consists of all ξ such that the linear prevision $\mathrm{Ef}^{\mathsf{a},\tau}(\bullet|\xi)$ can be normalized, i.e., such that $\mathsf{b}\xi > 0$. This choice results in a so-called full family and makes Ξ convex: take ξ' and ξ'' in Ξ and let $\lambda :]0,1[$, then

$$\begin{aligned}
\mathsf{b}(\lambda\cdot\xi' + (1-\lambda)\cdot\xi'') &= 1/\textstyle\sum_{z:\mathscr{X}} \exp\langle\lambda\cdot\xi' + (1-\lambda)\cdot\xi''|\tau z\rangle\cdot\mathsf{a}z \\
&= 1/\textstyle\sum_{z:\mathscr{X}} \big(\exp\langle\xi'|\tau z\rangle\cdot\mathsf{a}z\big)^{\lambda}\cdot\big(\exp\langle\xi''|\tau z\rangle\cdot\mathsf{a}z\big)^{1-\lambda} \\
&\geq \big(1/\textstyle\sum_{z:\mathscr{X}} \exp\langle\xi'|\tau z\rangle\cdot\mathsf{a}z\big)^{\lambda}\cdot\big(1/\textstyle\sum_{z:\mathscr{X}} \exp\langle\xi''|\tau z\rangle\cdot\mathsf{a}z\big)^{1-\lambda} \\
&= (\mathsf{b}\xi')^{\lambda}\cdot(\mathsf{b}\xi'')^{1-\lambda},
\end{aligned}$$

where Hölder's inequality [see, e.g., Schechter 1997, $\S 22.33_{593}$] was used in the third step; it follows that $\mathsf{b}(\lambda\cdot\xi' + (1-\lambda)\cdot\xi'') > 0$ and thus that $\lambda\cdot\xi' + (1-\lambda)\cdot\xi'' \in \Xi$. (We assumed $\mathscr{X}$ to be discrete; a completely similar argument works when $\mathscr{X}$ is continuous.)

(vii) We restrict our attention to exponential families that are regular, i.e., full families for which Ξ is an open set. This technical

restriction is necessary for a result from the literature (which we introduce in $(4.29)_{168}$).

(viii) From our point of view, the conceptually most important function is τ. It translates an observation z to a d-dimensional real vector τz in $\mathscr{T}$ and the way in which this is done betrays what is and what is not considered important in the sample. Whenever $\tau_{\mathscr{X}} = \mathrm{id}_{\mathscr{X}}$, the exponential family is called linear. A linear canonical exponential family is called natural [Letac 1992; Kotz et al. 2000, Ch. 54_{659}].

The function τ as a translation of what is and is not important in a sample, together with the structure of exponential families described by $(4.1)_{156}$, and the assumption of conditional independence determines the sufficient statistic associated with an exponential family.

To see how this works, we consider an observed sample sequence $\check{x}$ in $\mathscr{X}^*$ of length $\nu\check{x}$ (cf. §3.1.2$_{99}$). Due to the assumption of conditional independence, the corresponding likelihood is the product of the likelihoods for each of the samples in the sequence:

$$\begin{aligned} E_{\check{x}}^{\mathsf{a},\tau,\psi}\phi &:= \textstyle\prod_{k:1..\nu\check{x}} \mathsf{b}(\psi\phi)\cdot\exp\langle\psi\phi|\tau\check{x}_k\rangle\cdot\mathsf{a}\check{x}_k \\ &= \big(\mathsf{b}(\psi\phi)\big)^{\nu\check{x}}\cdot\exp\big\langle\psi\phi\,\big|\,\textstyle\sum_{k:1..\nu\check{x}}\tau\check{x}_k\big\rangle\cdot\prod_{k:1..\nu\check{x}}\mathsf{a}\check{x}_k \\ &= \big(\mathsf{b}(\psi\phi)\big)^{\nu\check{x}}\cdot\exp^{\nu\check{x}}\langle\psi\phi|\tau\check{x}\rangle\cdot\mathsf{a}\check{x}, \end{aligned} \tag{4.4}$$

where for any sample sequence x in $\mathscr{X}^*$, we have used the specifications $\tau x := \frac{1}{\nu x}\sum_{k:1..\nu x}\tau x_k$ and $\mathsf{a}x := \prod_{k:1..\nu x}\mathsf{a}x_k$, so these functions are fully defined by their single-sample case. This product shows that the order of the samples in a sequence is considered irrelevant: it is an ancillary statistic.

In this chapter, we restrict the scope of possible applications by assuming that the continuous likelihood principle holds (cf. §3.1.6$_{115}$). From this principle, it follows that the proportionality constant $\mathsf{a}\check{x}$ present in (4.4) is irrelevant for updating, which means that $(\nu\check{x},\tau\check{x})$ is the $(d+1)$-dimensional sufficient statistic: it then contains all the information in $\check{x}$ that under the given assumptions is deemed relevant for inference. When considering sample sequences of arbitrary length, we see that the range $\mathscr{T}$ of τ is the convex hull $\mathrm{co}\{\tau_{\mathscr{X}}\}$ of $\{\tau_{\mathscr{X}}\} = \{\tau z \,|\, z : \mathscr{X}\}$.

We have already mentioned that one of the most important properties of exponential family likelihood functions was their finite dimensional sufficient statistic. We see that for exponential families it consists of the number of observed samples and a d-dimensional real vector that can more-or-less be seen as an (arithmetic) mean single-sample sufficient statistic. It is instructive to see how the sufficient statistic of two observed sample sequences $\check{x}'$ and $\check{x}''$ are combined: their total length is $\nu(\check{x}',\check{x}'') = \nu\check{x}' + \nu\check{x}''$ and the new mean single-sample sufficient statistic $\tau(\check{x}',\check{x}'')$ becomes the convex mixture $\frac{\nu\check{x}'}{\nu(\check{x}',\check{x}'')}\cdot\tau\check{x}' + \frac{\nu\check{x}''}{\nu(\check{x}',\check{x}'')}\cdot\tau\check{x}''$.

The exponential family linear previsions for sequences in $\mathscr{Y} :\subset \mathscr{X}^*$

that correspond to the likelihood function (4.4) are (let f be a gamble in $\check{\mathscr{L}}_{\mathscr{Y}}$)

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(f|\mathscr{Y},\phi) = \sum_{y:\mathscr{Y}} f\,y \cdot E_y^{\mathsf{a},\tau,\psi}\phi, \quad \text{or} \tag{4.5}$$

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(f|\mathscr{Y},\phi) = \int_{\mathscr{Y}} f\,y \cdot E_y^{\mathsf{a},\tau,\psi}\phi \,\mathrm{d}y. \tag{4.6}$$

Here, $\mathscr{Y}$ is appropriately chosen in the sense that $\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\mathscr{Y}|\mathscr{Y},\phi) = 1$; for example, $\mathscr{Y}$ could consist of the set of sequences of some given constant length.

To finish this subsection, we direct our attention back to the single sample case $\mathscr{Y} = \mathscr{X}$ (cf. $(4.2)_{156}$ and $(4.3)_{156}$) and to the normalization function b. This function plays a central role in an interesting property of exponential family distributions: $\kappa := -\ln\circ\mathsf{b}$ is the so-called cumulant function of the exponential family distribution. Roughly, this is the logarithm of the moment generating function of the distribution. For us, the most interesting consequence – as we will see later – is that, vectorially

About cumulants, consult, e.g., Cramér [1946, §15.10$_{185}$].

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\tau|\phi) = (\nabla\kappa)(\psi\phi); \tag{4.7}$$

The nabla-symbol ∇ denotes the gradient operator.

component number $i : 1..d$ of this vectorial expression is

$$\mathrm{Ef}^{\mathsf{a},\tau,\psi}(\tau_i|\phi) = (\mathrm{D}_i\kappa)(\psi\phi), \tag{4.8}$$

Illustrating the partial derivative operator: let $f := u : \mathbb{R}^2 ; u_1 \cdot u_2^2$, then $(\mathrm{D}_2 f)(u_1,u_2) = 2 \cdot u_1 \cdot u_2$.

which is always defined, even if τ_i is unbounded [Barndorff-Nielsen 1978, Thm 8.1_{114}].

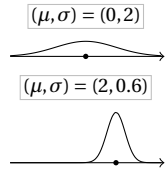
To make things more concrete, we give practical examples of exponential families in the next two subsections. The first, the normal family, is an example of a family defined on a continuous sample space and the second, the multivariate Bernoulli family, is an example of one that is defined on a discrete sample space. A lot of other commonly used sampling models such as exponential sampling and Poisson sampling also correspond to exponential families; for these two and also some others, a description in the same vein as the ones below can be found in the Bestiarium$_{202}$.

4.1.2 *Continuous example: normal sampling*

Our first example is the family of univariate normal distributions [see, e.g., Bernardo & Smith 1994, §3.2.2$_{121}$]; each member has $\mathscr{X} := \mathbb{R}$ as a sample space and is parameterized by its mean and standard deviation, i.e., a $\phi := (\mu,\sigma)$ in $\Phi := \mathbb{R}\times\mathbb{R}_{>0}$. Let f be any measurable gamble on $\mathbb{R}$; the normal linear prevision is then defined by

$$\mathrm{Nl}(f|\mu,\sigma) := \int_{\mathscr{X}} f\,z \cdot \tfrac{1}{\sqrt{2\cdot\pi}\cdot\sigma} \cdot \exp\Big(-\tfrac{1}{2}\cdot\big(\tfrac{z-\mu}{\sigma}\big)^2\Big)\,\mathrm{d}z. \tag{4.9}$$

Two normal densities (plot restricted to $[-5,5]$): (the mean μ is indicated with a dot)

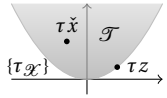


We start by rewriting the corresponding likelihood function in exponential family form: let z be an observed real sample, then this likelihood

can be written as

$$\tfrac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\Big(-\tfrac{1}{2}\cdot\big(\tfrac{z-\mu}{\sigma}\big)^2\Big)=\tfrac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\big(-\tfrac{z^2}{2\cdot\sigma^2}+\tfrac{z\cdot\mu}{\sigma^2}-\tfrac{\mu^2}{2\cdot\sigma^2}\big)$$
$$=\exp\big(-\tfrac{\mu^2}{2\cdot\sigma^2}-\ln\sigma\big)\cdot\exp\big\langle\tfrac{\mu}{\sigma^2},-\tfrac{1}{2\cdot\sigma^2}\,\big|\,z,z^2\big\rangle\cdot\tfrac{1}{\sqrt{2\cdot\pi}}.$$

Take $z:=0.4$ and $\check{x}:=(-1.1,0.5,-0.2)$:



Comparing this last expression with $(4.1)_{156}$, we see that the Euclidean vector space for this family has dimension $d:=2$ and that

$$\begin{aligned}&\tau:=z:\mathbb{R};(z,z^2),\quad\text{so}\quad\mathscr{T}:=\mathrm{co}\{z,z^2\,|\,z:\mathbb{R}\}=\{t:\mathbb{R}^2\,|\,t_2\geq t_1^2\},\\&\psi:=(\mu,\sigma):\mathbb{R}\times\mathbb{R}_{>0};(\tfrac{\mu}{\sigma^2},-\tfrac{1}{2\cdot\sigma^2}),\quad\text{so}\quad\Xi:=\mathbb{R}\times\mathbb{R}_{<0},\end{aligned}\tag{4.10}$$

and also that

$$\mathsf{a}:=\mathbb{R};\tfrac{1}{\sqrt{2\cdot\pi}}\quad\text{and}\quad\mathsf{b}:=\xi:\mathbb{R}\times\mathbb{R}_{<0};\exp\Big(\tfrac{1}{2}\cdot\big(\tfrac{\xi_1^2}{2\cdot\xi_2}+\ln(-2\cdot\xi_2)\big)\Big).\tag{4.11}$$

In principle, constant factors (such as $-1/2$) can be swapped between τ and ψ. Our choice for τ was inspired by a possible assumption that justifies using a normal likelihood function: that any sample sequence $\check{x}:\mathscr{X}^*$ with the same sequence length $\nu\check{x}$, the same barycenter $\tau_1\check{x}$, and the same moment of inertia $\tau_2\check{x}-(\tau_1\check{x})^2$ is equally likely to occur.

The cumulant function for the normal family becomes

$$\kappa:=-\ln\circ\mathsf{b}=\xi:\mathbb{R}\times\mathbb{R}_{<0};-\tfrac{1}{2}\cdot\big(\tfrac{\xi_1^2}{2\cdot\xi_2}+\ln(-2\cdot\xi_2)\big).$$

So, using $(4.7)_{\frown}$ and $(4.8)_{\frown}$, we find – letting $\xi:=\psi(\mu,\sigma)$ – that

$$\mathrm{Nl}(\tau|\mu,\sigma)=(\nabla\kappa)\xi=-\tfrac{1}{2}\cdot\big(\tfrac{\xi_1}{\xi_2},-\tfrac{\xi_1^2}{2\cdot\xi_2^2}+\tfrac{1}{\xi_2}\big)=(\mu,\mu^2+\sigma^2),\tag{4.12}$$

i.e., a vector consisting of the mean and the second-order noncentral moment.

4.1.3 *Discrete example: multi-category Bernoulli & negative multinomial sampling*

We start our second example by investigating the multi-category Bernoulli family of distributions. The archetypical example of a Bernoulli trial is the drawing of one marble with replacement from a bag of variously colored marbles. The repetition of Bernoulli trials lies at the basis of the already quite familiar multinomial sampling process, but also of others, e.g., the negative multinomial sampling process.

Each multi-category Bernoulli distribution has a nonempty finite category set $\mathscr{X}$ as a sample space and is parameterized by a $|\mathscr{X}|$-dimensional frequency vector $\phi:=\vartheta$ from the $(|\mathscr{X}|-1)$-dimensional open unit simplex $\Phi:=\mathrm{int}\,\Delta_{\mathscr{X}}$. So strictly speaking, there is a Bernoulli family for each possible finite set $\mathscr{X}$. Luckily, thanks to our capacity for abstract thought, we can consider $\mathscr{X}$ generically and treat all these families in one fell swoop.

Illustrating the interior operator: $\mathrm{int}[0,1]=]0,1[$.

Let f be a gamble on $\mathscr{X}$, the multi-category Bernoulli linear prevision is then defined by

$$\mathrm{Br}(f|\vartheta) := \textstyle\sum_{z:\mathscr{X}} f z \cdot \vartheta_z = \mathrm{Mn}(f|1,\vartheta) = \mathrm{Wa}(f|\vartheta). \tag{4.13}$$

Now let z be an observed category, the corresponding Bernoulli likelihood is then simply

$$\vartheta_z = \textstyle\prod_{z':\mathscr{X}} \vartheta_{z'}^{(C_{\mathscr{X}} z)_{z'}} = \exp\big(\sum_{z':\mathscr{X}} \ln \vartheta_{z'} \cdot (C_{\mathscr{X}} z)_{z'}\big),$$

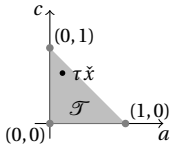
where $C_{\mathscr{X}}$ is the counting map $(3.9)_{100}$ from the previous chapter (also recall the count vector space notation $(3.8)_{99}$). This expression is well-defined because $\vartheta > 0$, being a member of the *open* unit simplex. The last expression already seems to be in the form of $(4.1)_{156}$, but written as such, it would not correspond to a minimal exponential family: the counts for the different categories are not affinely independent, i.e., $\sum(C_{\mathscr{X}} z) = 1$. Therefore, we single out some (any) category $o : \mathscr{X}$ to allow us to eliminate this dependence from the likelihood's expression. We know that $(C_{\mathscr{X}} z)_o = 1 - \sum_{z':\mathscr{X}_{\neq o}} (C_{\mathscr{X}} z)_{z'}$, so

$$\begin{aligned}
\vartheta_z &= \exp\Big(\textstyle\sum_{z':\mathscr{X}_{\neq o}} \ln \vartheta_{z'} \cdot (C_{\mathscr{X}} z)_{z'} + \ln \vartheta_o \cdot \big(1 - \sum_{z':\mathscr{X}_{\neq o}} (C_{\mathscr{X}} z)_{z'}\big)\Big) \\
&= \exp\big(\textstyle\sum_{z':\mathscr{X}_{\neq o}} \ln \frac{\vartheta_{z'}}{\vartheta_o} \cdot (C_{\mathscr{X}} z)_{z'} + \ln \vartheta_o\big) \\
&= \vartheta_o \cdot \exp\Big\langle \big(\ln \tfrac{\vartheta_{z'}}{\vartheta_o} \,\big|\, z' : \mathscr{X}_{\neq o}\big) \,\Big|\, (C_{\mathscr{X}} z)_{\mathscr{X}_{\neq o}} \Big\rangle.
\end{aligned} \tag{4.14}$$

Comparing this last expression to $(4.1)_{156}$, we see that the Euclidean vector space for this family has dimension $d := |\mathscr{X}_{\neq o}|$ and that

$$\begin{aligned}
\tau := z : \mathscr{X} ; (C_{\mathscr{X}} z)_{\mathscr{X}_{\neq o}}, \quad \text{so} \quad \mathscr{T} &:= \mathrm{co}\big(\iota \mathscr{X}_{\neq o} ; 0 \cup 1^{\mathscr{X}_{\neq o}}\big) \\
&= \big\{ t : (\mathbb{R}_{\geq 0})^{\mathscr{X}_{\neq o}} \,\big|\, \textstyle\sum t \leq 1 \big\}, \\
\psi := \vartheta : \mathrm{int}\, \Delta_{\mathscr{X}} ; \big(\ln \tfrac{\vartheta_{z'}}{\vartheta_o} \,\big|\, z' : \mathscr{X}_{\neq o}\big), \quad \text{so} \quad \Xi &:= \mathbb{R}^{\mathscr{X}_{\neq o}},
\end{aligned} \tag{4.15}$$

Take $\mathscr{X} := \{a, b, c\}$, $\check{x} := cbcacc$, and let $o := b$:



where we see that restricting the frequency vectors to the open simplex was necessary to guarantee the openness of Ξ. We also see, using the identity $\vartheta_o = 1 - \sum_{z':\mathscr{X}_{\neq o}} \vartheta_{z'}$, that

$$\mathsf{a} := \mathscr{X} ; 1 \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R}^{\mathscr{X}_{\neq o}} ; \frac{1}{1 + \sum_{z':\mathscr{X}_{\neq o}} \exp \xi_{z'}}. \tag{4.16}$$

The Bernoulli linear prevision $\mathrm{Br}(\bullet|\vartheta)$ is fully defined by these functions and by $(4.2)_{156}$.

Note that there is, again, with the choice of the elimination category o, some freedom in the way exponential family likelihoods can be written in their exponential family form. One should not fear that this reformulation places this elimination category in a special position: as ψ is invertible, we can always rewrite an expression using canonical parameters into an elimination-category agnostic form. The reformulation

in exponential family form is only meant to bring to light the common mathematical structure of these families. It is this common structure which is later on exploited to introduce structurally similar imprecise-probabilistic inference models for all these families at the same time; it is usually preferable from the interpretational standpoint to use the original, classical parameterization and form.

The sufficient statistic corresponding to an observed sample sequence $\check{x}$ in $\mathscr{X}^*$ is $(\nu\check{x}, \tau\check{x}) = \big(\nu\check{x}, (C_{\mathscr{X}}\check{x})_{\mathscr{X}_{\neq o}}/\nu\check{x}\big)$, or, rewriting the information in a more familiar way, the count vector $C_{\mathscr{X}}\check{x}$ (cf. §3.1.5$_{112}$). We already know that an assumption of exchangeability is enough to justify this sufficient statistic and thus the τ used.

The cumulant function for the $\mathscr{X}$-Bernoulli family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi : \mathbb{R}^{\mathscr{X}_{\neq o}} ; -\ln\big(1+\textstyle\sum_{z':\mathscr{X}_{\neq o}} \exp\xi_{z'}\big).$$

So, using (4.7)$_{159}$ and (4.8)$_{159}$, we find – letting $\xi := \psi\vartheta$ – that

$$\mathrm{Br}(\tau|\vartheta) = (\nabla\kappa)\xi = \left(\frac{\exp\xi_{z'}}{1+\sum_{z':\mathscr{X}_{\neq o}}\exp\xi_{z'}} \,\middle|\, z':\mathscr{X}_{\neq o}\right) = (\vartheta_{z'} \mid z':\mathscr{X}_{\neq o}), \quad (4.17)$$

i.e., grosso modo the expectation of the frequency vector τ is the family member's defining frequency vector ϑ itself.

For purposes of predictive inference, it is necessary to know the linear prevision for the type of sequences we wish to make predictions about. We have already treated the problem of making predictions about sequences of fixed length extensively in §3.2$_{118}$. The predictive inference model we used was the IDMM, which is related to its parametric cousin, the IDM, through (count) multinomial previsions (cf. §3.3.1$_{139}$). The multinomial prevision $\mathrm{Mn}(\circledast|n,\vartheta)$ for sequences of length $n:\mathbb{N}$, originally defined by (3.33)$_{107}$, corresponds to the generalization of the Bernoulli prevision via (4.5)$_{159}$ for the set of sequences $\mathscr{Y} := \mathscr{X}^n$.

Other interesting types of sequences can be investigated, e.g., those that stop after a pre-specified subsequence has been observed or those that stop after a pre-specified minimum number of observations from each category has been seen. We quickly work out the latter example: given a count vector m in $\mathbb{N}^{\mathscr{X}}$, we wish to describe the relative likelihoods of all the sequences with a count vector that only just dominates m in the sense that the last sample in the sequence is necessary to guarantee this dominance. So we take

$$\mathscr{Y} := \big\{x : \mathscr{X}^* \,\big|\, C_{\mathscr{X}}x \geq m \wedge (C_{\mathscr{X}}x)_{x_{\nu x}} = m_{x_{\nu x}}\big\}.$$

Using this set and (4.5)$_{159}$, we can write down the expression for the corresponding negative multinomial prevision: (let f be a gamble on $\mathscr{Y}$)

$$\mathrm{Nm}(f|m,\vartheta) := \sum_{\substack{x:\mathscr{X}^* \wedge C_{\mathscr{X}}x\geq m \\ \wedge (C_{\mathscr{X}}x)_{x_{\nu x}} = m_{x_{\nu x}}}} fx \cdot \textstyle\prod_{z:\mathscr{X}} \vartheta_z^{(C_{\mathscr{X}}x)_z}; \quad (4.18)$$

as the last observed sample x_{vx} plays a distinct role, we separate it:

$$= \sum_{z:\mathscr{X}_m} \sum_{y:\mathscr{X}^* \wedge C_{\mathscr{X}}(y,z) \geq m \atop \wedge (C_{\mathscr{X}} y)_z = m_z - 1} f(y,z) \cdot \vartheta_z \cdot \prod_{z':\mathscr{X}} \vartheta_{z'}^{(C_{\mathscr{X}} y)_{z'}};$$

we can group things per atom and replace the last factor by a Bernstein polynomial $(3.34)_{107}$:

$$= \sum_{z:\mathscr{X}_m} \vartheta_z \cdot \sum_{m':\mathbb{N}^{\mathscr{X}} \wedge m' \geq m - C_{\mathscr{X}} z \atop \wedge m'_z = m_z - 1} \sum_{y:[m']} f(y,z) \cdot B_y \vartheta;$$

and finally, using $(3.25)_{104}$, recognize the multivariate hypergeometric prevision $(3.18)_{101}$:

$$= \sum_{z:\mathscr{X}_m} \vartheta_z \cdot \sum_{m':\mathbb{N}^{\mathscr{X}} \wedge m' \geq m - C_{\mathscr{X}} z \atop \wedge m'_z = m_z - 1} \mathrm{Mh}\big(f(\circledcirc, z) \;\big|\; m'\big) \cdot B_{m'} \vartheta.$$

We have worked towards this last expression, because it allows us to introduce – inspired by the analogy between the sequence and count-multinomial (cf. $(3.31)_{106}$) – the negative count-multinomial prevision (let h be a gamble on $\{m' : \mathbb{N}^{\mathscr{X}} \mid m' \geq m \wedge \exists z : \mathscr{X} ; m'_z = m_z\}$)

$$\mathrm{Cn}(h|m,\vartheta) := \sum_{z:\mathscr{X}_m} \vartheta_z \cdot \sum_{m':\mathbb{N}^{\mathscr{X}} \wedge m' \geq m - C_{\mathscr{X}} z \atop \wedge m'_z = m_z - 1} h(m' + C_{\mathscr{X}} z) \cdot B_{m'} \vartheta \quad (4.19)$$

(To my surprise, this definition turns out to be more general than the one typically found in the literature [Johnson et al. 1997, §36.1$_{93}$], which treats the case of m that are nonzero only in one category.)

We have not only derived the expression (4.19) for the negative count-multinomial distribution just because we could. It also allows us to make a number of remarks about the corresponding likelihood function: assume we have observed a negative multinomial sample sequence that ended with a sample z and was reported to further consist of a count vector $\check{m} : \mathbb{N}^{\mathscr{X}} \wedge \check{m} \geq m \wedge \exists z : \mathscr{X} ; \check{m}_z = m_z$, then

$$\begin{aligned} \vartheta_z \cdot B_{\check{m}} \vartheta &= |[\check{m}]| \cdot \prod_{z':\mathscr{X}} \vartheta_{z'}^{(\check{m} + C_{\mathscr{X}} z)_{z'}} \\ &= (\vartheta_o)^{v\check{m}+1} \cdot \exp^{v\check{m}+1} \Big\langle \big(\ln \tfrac{\vartheta_{z'}}{\vartheta_o} \;\big|\; z' : \mathscr{X}_{\neq o}\big) \;\Big|\; \big(\tfrac{\check{m} + C_{\mathscr{X}} z}{v\check{m}+1}\big)_{\mathscr{X}_{\neq o}} \Big\rangle \cdot |[\check{m}]|, \end{aligned}$$

where the last step summarizes a derivation similar to the one leading to $(4.14)_{161}$. Not unexpectedly, this expression looks very much like the one for sequences $(4.4)_{158}$ of Bernoulli samples of length $v\check{m}+1$ with sufficient statistic $\check{m} + C_{\mathscr{X}} z$. It does not fit into that framework entirely, due to the transition from sequences to counts. Namely, there is an extra proportionality factor $|[\check{m}]|$, the number of sequences compatible with the observation – which, because of the likelihood principle, plays no role in updating –, and of course the way in which the observation is translated into a mean single-sample sufficient statistic has changed. This illustrates that apart from the distributions with likelihoods that

can be put in the classical exponential family form for single samples or sequences ($(4.1)_{156}$ and $(4.4)_{158}$, respectively), we can also consider those that are in some way derived from them.

The extensive examples in this and the previous subsection close our discussion of exponential family likelihoods and previsions. We next continue with their friends.

4.1.4 *Conjugate linear previsions*

The friends we refer to, are classes of parametric and predictive linear previsions, one for each exponential family. They will constitute the basic building blocks for the imprecise-probabilistic inference models for exponential families that form the subject of this chapter.

We have seen in $\S 3.1.6_{115}$ that in a classical Bayesian context, updating is done by combining a prior linear prevision with the likelihood function for the observation using Bayes's rule. Or, when using exponential family likelihoods, Bayes's rule for density functions [see, e.g., Walley 1991, $\S 6.10.4_{331}$]. The reason for this is that the prior prevision is then defined on a continuous possibility space, the set of parameters Φ or its image under ψ, the convex set of canonical parameters Ξ (cf. $(\text{vi})_{157}$), and that we here only consider prior linear previsions that can be written using density functions (i.e., that are – technically speaking – σ-finite additive functions of the Lebesgue measurable subsets of Ξ that are absolutely continuous with respect to Lebesgue measure on Ξ).

Let R be some linear prevision on the set of all measurable gambles $\mathscr{L}_\Xi$ defined by the probability density function $r : \Xi \to \mathbb{R}_{\geq 0}$ (so with $\int_\Xi r = 1$); i.e., if f is a gamble in $\mathscr{L}_\Xi$, then

$$Rf = \int_\Xi (f \cdot r).$$

If we use this prevision as a prior for our uncertainty about the parameter of a sampling process that we assume to be distributed according to a regular (canonical) exponential family with defining functions a and τ, then, after observing a sample sequence $\check{x}$, the posterior linear prevision $R(\bullet|\nu\check{x}, \tau\check{x})$ on $\mathscr{L}_\Phi$ is defined by

$$R(f|\nu\check{x}, \tau\check{x}) = \tfrac{1}{RE_{\check{x}}^{\mathsf{a},\tau}} \cdot R(f \cdot E_{\check{x}}^{\mathsf{a},\tau}), \tag{4.20}$$

which follows from Bayes's rule for density functions; the updated density is

$$\tfrac{1}{RE_{\check{x}}^{\mathsf{a},\tau}} \cdot r \cdot E_{\check{x}}^{\mathsf{a},\tau}.$$

Of course this only holds whenever $RE_{\check{x}}^{\mathsf{a},\tau} > 0$; otherwise, if $RE_{\check{x}}^{\mathsf{a},\tau} = 0$, the updated prevision is vacuous (cf. the for linear previsions identical $(1.83)_{61}$ and $(1.70)_{58}$). Note that we represent the conditioning event using the sufficient statistic and not the observed sequence itself; this

was done to stress that all other information in the observation is ancillary: irrelevant for inference purposes.

Out of considerations of mathematical tractability, we are going to limit ourselves to conjugate priors (cf. §3.3.2$_{140}$). Let us recall that, roughly speaking, this means that the prior and posterior uncertainty models have the same functional form. The prior and posterior uncertainty models we consider here are completely determined by their probability density functions. In the previous paragraph we have seen that the posterior density function is proportional to the product of the prior density function and the likelihood function. So the probability density functions of the conjugate family of some exponential family must have a form that remains essentially unchanged after multiplication by the exponential family likelihood (4.4)$_{158}$. This consideration leads to the following proposed expression for a linear prevision that could be used as a prior parametric inference model for the exponential family defined by a and τ: (f is still a measurable gamble on Ξ)

$$\mathrm{Cf}^{\mathsf{a},\tau}(f|s,t) := \int_{\Xi} f\xi \cdot \mathsf{c}(s,t) \cdot (\mathsf{b}\xi)^s \cdot \exp^s\langle \xi | t\rangle \,\mathrm{d}\xi, \tag{4.21}$$

where the possible values of the so-called hyperparameters s and t are limited to $\mathbb{R}_{>0}$ and $\operatorname{int}\mathscr{T}$ to guarantee normalizability [Diaconis & Ylvisaker 1979, Thm 1] and allow for a nice interpretation. The function $\mathsf{c}:\mathbb{R}_{>0}\times\operatorname{int}\mathscr{T}\to\mathbb{R}_{>0}$ provides the normalization factor; it is defined by

$$\mathsf{c}(s,t) = 1/\int_{\Xi} (\mathsf{b}\xi)^s \cdot \exp^s\langle \xi | t\rangle \,\mathrm{d}\xi. \tag{4.22}$$

When working with noncanonical parameters – belonging to a set Φ related to Ξ by a function $\psi:\Phi\to\Xi$ –, we can derive a conjugate prevision defined on $\check{\mathscr{L}}_{\Phi}$ from (4.21) [Burrill 1972, Thm 8-7$_{163}$]: (let g be a measurable gamble on Φ)

$$\mathrm{Cf}^{\mathsf{a},\tau,\psi}(g|s,t) := \int_{\Phi} g\phi \cdot \mathsf{c}(s,t) \cdot \big(\mathsf{b}(\psi\phi)\big)^s \cdot \exp^s\langle \psi\phi | t\rangle \cdot \big|\big|(\nabla\psi)\phi\big|\big| \,\mathrm{d}\phi, \tag{4.23}$$

About Jacobians and restrictions on ψ, consult, e.g., Cramér [1946, §22.2$_{291}$].

where $\nabla\psi$ stands for the transformation's Jacobian function (the gradient is applied componentwise to the vector function ψ). We have not defined the conjugate prevision directly in terms of noncanonical parameters (i.e., the above expression, but *without* the Jacobian and with an appropriately modified normalization factor), because in general it does not possess the interesting posterior weighted average prediction property (4.29)$_{168}$ [Gutiérrez-Peña & Smith 1995], which we will encounter later.

Arnold et al. [1993] give the most general conjugate families; Kotz et al. [2000, §54.7.3$_{3676}$] and Gutiérrez-Peña & Smith [1997] give an overview of the relevant literature.

The nice interpretation we can attach to the two hyperparameters becomes apparent when we update the prior (4.21) after the observation of a sample sequence $\check{x}$. Following (4.20), we need only squeeze (4.4)$_{158}$

under the integral sign to see that – as intended – the posterior linear prevision belongs to the same family:

$$\begin{aligned}\mathrm{Cf}^{\mathsf{a},\tau}(f|s,t,\nu\check{x},\tau\check{x}) &= \mathrm{Cf}^{\mathsf{a},\tau}\Big(f \;\Big|\; s+\nu\check{x}, \tfrac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}\Big)\\ &= \int_{\Xi} f\xi\cdot \mathsf{c}\big(s+\nu\check{x}, \tfrac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}\big)\cdot(\mathsf{b}\xi)^{s+\nu\check{x}} \\ &\qquad \cdot \exp^{s+\nu\check{x}}\big\langle \xi \;\big|\; \tfrac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}\big\rangle \,\mathrm{d}\xi\end{aligned} \tag{4.24}$$

has the same form as the prior $(4.21)_{\cap}$, but with updated hyperparameters $s+\nu\check{x}$ in $\mathbb{R}_{>0}$ and $\frac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}} = \frac{s}{s+\nu\check{x}}\cdot t + \frac{\nu\check{x}}{s+\nu\check{x}}\cdot\tau\check{x}$ in $\mathrm{int}\,\mathscr{T}$. As both $E^{\mathsf{a},\tau}_{\check{x}}$ and the prior conjugate prevision's density function are strictly positive, $\mathrm{Cf}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,t) > 0$, so the posterior prevision is never vacuous.

We see that s can be interpreted as a hypothetical number of counts that gives the weight of the prior assessment; it is called the number of pseudocounts. Similarly, t can be seen as a hypothetical mean single-sample sufficient statistic which we call the pseudomean. Due to the additivity of counts and the taking of weighted averages of mean single-sample sufficient statistics in the updating procedure, it does not matter for the final result if the updating is done in a step-by-step fashion – one sample (sequence) at a time – or in batch – i.e., in one updating run, after lumping all samples together in one big sequence (also see the end of the paragraph before $(4.5)_{159}$).

The posterior (4.24) can be rewritten in terms of a noncanonical parameter in a completely analogous way to how $(4.23)_{\cap}$ was derived from $(4.21)_{\cap}$. Also, in this section we are not going to give explicit posterior expressions anymore, as they can be derived from the prior expression by substituting $s+\nu\check{x}$ for s and $\frac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}$ for t.

4.1.5 *Predictive linear previsions*

A typical use of the conjugate parametric priors and posteriors introduced in the previous subsection is as a basis for predictive inference: Consider the case of a subject who wishes to make predictions about still unobserved samples (e.g., future samples). She knows or assumes that the samples are drawn from an unknown member of a known exponential family of distributions. To model her uncertainty about the member – through its defining parameter –, she uses a prior or posterior conjugate prevision of the type we have just seen. Such a conjugate prevision can be used as a second-order model, the first-order model being the exponential family prevision with unknown parameter. So, given the exponential family-based sampling model characterized by the set $\mathscr{Y} :\subset \mathscr{X}^*$ of possible unobserved sample sequences (cf. (4.5)–$(4.6)_{159}$), the corresponding prior predictive linear prevision $\mathrm{Pf}^{\mathsf{a},\tau}(\bullet|\mathscr{Y},s,t)$ on $\mathscr{L}_{\mathscr{Y}}$ is defined by (let f be a gamble in $\check{\mathscr{L}}_{\mathscr{Y}}$)

$$\mathrm{Pf}^{\mathsf{a},\tau}(f|\mathscr{Y},s,t) = \mathrm{Cf}^{\mathsf{a},\tau}\big(\mathrm{Ef}^{\mathsf{a},\tau}(f|\mathscr{Y},\bullet) \;\big|\; s,t\big); \tag{4.25}$$

which becomes, using the prior's expression $(4.21)_{165}$ and the exponential family prevision's expression $(4.6)_{159}$ for continuous sample spaces:

$$= \int_{\Xi} \Big(\int_{\mathscr{Y}} f\,y \cdot E_y^{\mathsf{a},\tau}\,\xi \,\mathrm{d}y \Big) \cdot \mathsf{c}(s,t) \cdot (\mathsf{b}\xi)^s \cdot \exp^s \langle \xi | t \rangle \,\mathrm{d}\xi;$$

writing out the likelihood functions's expression (4.4) and using Fubini's theorem to switch the order of integration gives us

$$= \int_{\mathscr{Y}} f\,y \cdot \mathsf{a}y \cdot \mathsf{c}(s,t) \cdot \Big(\int_{\Xi} (\mathsf{b}\xi)^{s+\nu y} \cdot \exp^{s+\nu y} \big\langle \xi \,\big|\, \tfrac{s\cdot t + \nu y \cdot \tau y}{s+\nu y} \big\rangle \,\mathrm{d}\xi \Big) \,\mathrm{d}y.$$

Then $(4.22)_{165}$ allows us to write this as

$$\mathrm{Pf}^{\mathsf{a},\tau}(f|\mathscr{Y},s,t) = \int_{\mathscr{Y}} f\,y \cdot \frac{\mathsf{a}y \cdot \mathsf{c}(s,t)}{\mathsf{c}\big(s+\nu y, \frac{s\cdot t + \nu y \cdot \tau y}{s+\nu y}\big)} \,\mathrm{d}y. \tag{4.26}$$

An entirely similar derivation holds for discrete sample spaces: in the first step we use $(4.5)_{159}$ instead of $(4.6)_{159}$, the rest of the steps stay essentially the same; this results in the defining expression

$$\mathrm{Pf}^{\mathsf{a},\tau}(f|\mathscr{Y},s,t) = \sum_{y:\mathscr{Y}} f\,y \cdot \frac{\mathsf{a}y \cdot \mathsf{c}(s,t)}{\mathsf{c}\big(s+\nu y, \frac{s\cdot t + \nu y \cdot \tau y}{s+\nu y}\big)}. \tag{4.27}$$

Remark that due to our choice of conjugate prevision, these predictive previsions are independent of the parameterization used for the exponential family. For immediate prediction – i.e., when $\mathscr{Y} = \mathscr{X}$ – we need not and do not mention $\mathscr{Y}$.

Note that the probability mass or density functions for both the discrete and continuous case have the same form. As we need it further on in §4.3.3$_{186}$, we give the explicit expression for the immediate prediction case: (let z be a potential sample from $\mathscr{X}$)

$$\mathrm{pf}^{\mathsf{a},\tau}(z|s,t) := \frac{\mathsf{a}z \cdot \mathsf{c}(s,t)}{\mathsf{c}\big(s+1, \frac{s\cdot t + \tau z}{s+1}\big)}. \tag{4.28}$$

As it is composed of strictly positive factors (see $(\mathrm{i})_{156}$ and $(4.22)_{165}$), this probability mass or density is strictly positive everywhere on $\mathscr{X}$.

Derivations similar to (4.25)–$(4.27)_{166–167}$, but now starting from a posterior conjugate prevision (4.24), lead to expressions for the posterior predictive linear previsions. As mentioned, they are identical to the prior ones above up to the substitution of $s+\nu\check{x}$ for s and $\frac{s\cdot t+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}$ for t. Therefore $s+\nu\check{x}+\nu y$ must be substituted for $s+\nu y$ and the weighted average $\frac{s\cdot t+\nu\check{x}\cdot\tau\check{x}+\nu y\cdot\tau y}{s+\nu\check{x}+\nu y}$ for $\frac{s\cdot t+\nu y\cdot\tau y}{s+\nu y}$.

To close this subsection, there is one interesting property of the conjugate and immediate predictive previsions we must mention. We have seen in $(4.7)_{159}$ that $\mathrm{Ef}^{\mathsf{a},\tau}(\tau|\bullet) = \nabla\kappa$, where, recall, κ is the cumulant function of the exponential family distribution. So then (4.25) tells us

that

$$\mathrm{Pf}^{\mathsf{a},\tau}(\tau|s,t) = \mathrm{Cf}^{\mathsf{a},\tau}(\nabla\kappa|s,t) = t, \tag{4.29}$$

where we used vectorial notation and where the second equality is a nontrivial result by Diaconis & Ylvisaker [1979, Thm 2]. We call this the posterior weighted average prediction property: the posterior prevision of a single-sample sufficient statistic is the weighted average $\frac{s}{s+v\check{x}}\cdot t + \frac{v\check{x}}{s+v\check{x}}\cdot \tau\check{x}$ of the pseudomean t and the observed mean single-sample sufficient statistic $\tau\check{x}$. Actually, this property *characterizes* the conjugate families we used for all regular exponential families on a continuous possibility space [Diaconis & Ylvisaker 1979, Thm 3] and for many common regular exponential families on a discrete possibility space [Diaconis & Ylvisaker 1979, Thms 4 & 5; Johnson 1967]: the members of these conjugate families are the only nondegenerate ones that possess the posterior weighted average prediction property.

It is now time to make the things discussed here more concrete with our two example exponential families.

4.1.6 *Conjugate & predictive linear previsions for normal sampling*

We first return to our example of normal sampling, first introduced in §4.1.2$_{159}$.

The information in (4.10)$_{160}$ and (4.11)$_{160}$ fully defines the canonical form (4.21)$_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}\times\mathbb{R}_{<0}$ and the domain of possible parameters is $\operatorname{int}\mathscr{T} = \{t : \mathbb{R}^2 \mid t_2 > t_1^2\}$. The one thing that we do not yet know explicitly is the normalization factor (4.22)$_{165}$: (let s be a strictly positive real number and $t : \operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_\Xi \exp^s\Big(\frac{1}{2}\cdot\big(\frac{\xi_1^2}{2\cdot\xi_2} + \ln(-2\cdot\xi_2)\big)\Big)\cdot\exp^s\langle\xi|t\rangle\,\mathrm{d}\xi.$$

It can be calculated more easily when starting from the noncanonical parameterization given by the function $\psi' := (\mu,\lambda) : \mathbb{R}\times\mathbb{R}_{>0}\,;(\mu\cdot\lambda, -\frac{1}{2}\cdot\lambda)$, where μ is the normal distribution's mean and λ its precision (related to the standard deviation σ by $\lambda = 1/\sigma^2$):

$$\begin{aligned}\mathsf{c}(s,t) = 1/\textstyle\int_{\mathbb{R}\times\mathbb{R}_{>0}} &\exp^s\big(\tfrac{1}{2}\cdot(-\mu^2\cdot\lambda + \ln\lambda) + \mu\cdot\lambda\cdot t_1 - \tfrac{1}{2}\cdot\lambda\cdot t_2\big)\\ &\cdot\big|\big|(\nabla\psi')(\mu,\lambda)\big|\big|\,\mathrm{d}\mu\,\mathrm{d}\lambda;\end{aligned}$$

applying the Fubini theorem [Burrill 1972, §7-6$_{125}$] gives

$$\begin{aligned}= 1/\textstyle\int_{\mathbb{R}_{>0}}\Big(\int_{\mathbb{R}}\lambda^{\frac{s}{2}}\cdot\exp^s\big(-\tfrac{1}{2}\cdot(\lambda\cdot\mu^2 - 2\cdot\lambda\cdot\mu\cdot t_1) - \tfrac{1}{2}\cdot\lambda\cdot t_2\big)\\ \cdot\Big|\begin{vmatrix}\lambda & 0\\ \mu & -\frac{1}{2}\end{vmatrix}\Big|\,\mathrm{d}\mu\Big)\,\mathrm{d}\lambda;\end{aligned}$$

completing the square of the exponent's first term gives

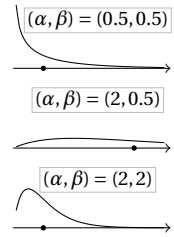
$$\begin{aligned}
&= 1/\textstyle\int_{\mathbb{R}_{>0}}\Big(\int_{\mathbb{R}}\lambda^{\frac{s}{2}}\cdot\exp^{s}\big(-\tfrac{\lambda}{2}\cdot(\mu-t_1)^2-\tfrac{\lambda}{2}\cdot(t_2-t_1^2)\big)\cdot\tfrac{1}{2}\cdot\lambda\,\mathrm{d}\mu\Big)\,\mathrm{d}\lambda\\
&= 2/\textstyle\int_{\mathbb{R}_{>0}}\Big(\int_{\mathbb{R}}\lambda^{\frac{1}{2}}\cdot\exp\big(-\tfrac{s\cdot\lambda}{2}\cdot(\mu-t_1)^2\big)\,\mathrm{d}\mu\Big)\\
&\qquad\cdot\lambda^{\frac{s+1}{2}}\cdot\exp\big(-\tfrac{s}{2}\cdot(t_2-t_1^2)\cdot\lambda\big)\,\mathrm{d}\lambda\\
&= 2\cdot\frac{\big(\frac{s}{2}\cdot(t_2-t_1^2)\big)^{\frac{s+3}{2}}}{\Gamma\frac{s+3}{2}}\cdot\sqrt{\tfrac{s}{2\cdot\pi}}\cdot
\end{aligned}\tag{4.30}$$

The last step follows from the fact that the inner integral is proportional to $\mathrm{Nl}\big(\mathbb{R}\;\big|\;t_1,1/\sqrt{s\cdot\lambda}\big)$ and therefore the entire expression is proportional to $\mathrm{Ga}\big(\mathbb{R}_{>0}\;\big|\;\frac{s+3}{2},\frac{s}{2}\cdot(t_2-t_1^2)\big)$, where (let $f:\mathscr{L}_{\mathbb{R}_{>0}}$)

$$\mathrm{Ga}(f|\alpha,\beta)\coloneqq\textstyle\int_{\mathbb{R}_{>0}}f\,r\cdot\frac{\beta^{\alpha}}{\Gamma\alpha}\cdot r^{\alpha-1}\cdot\exp(-\beta\cdot r)\,\mathrm{d}r\tag{4.31}$$

is the gamma linear prevision [see, e.g., Bernardo & Smith 1994, §3.2.2$_{118}$], with strictly positive real parameters α and β.

Some gamma densities (plot restricted to $]0,5]$): (mean α/β indicated with a dot)



For quite obvious reasons the conjugate family is called the normal-gamma family. The normal-gamma linear prevision [see, e.g., Bernardo & Smith 1994, §3.2.5$_{136}$] is defined by (now let f be a measurable gamble on the space $\mathbb{R}\times\mathbb{R}_{>0}$)

$$\mathrm{Ng}(f|\mu,\rho,\alpha,\beta)\coloneqq\mathrm{Ga}\big(\mathrm{Nl}(f|\mu,1/\sqrt{\rho\cdot\bullet})\;\big|\;\alpha,\beta\big),\tag{4.32}$$

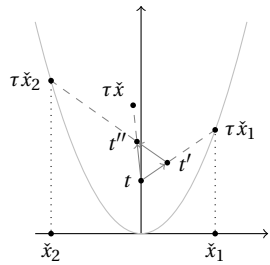
with parameters restricted as before: μ – not to be confused with earlier uses of this symbol – can be any real number, but ρ must be strictly positive. So the prior parametric linear prevision is

$$\mathrm{Ng}\big(\bullet\;\big|\;t_1,s,\tfrac{s+3}{2},\tfrac{s}{2}\cdot(t_2-t_1^2)\big).\tag{4.33}$$

(Note that not every normal-gamma prevision belongs to the conjugate family.) The posterior is obtained by substituting $s+\nu\check{x}$ for s, $\frac{s\cdot t_1+\sum\check{x}}{s+\nu\check{x}}$ for t_1, and $\frac{s\cdot t_2+\sum\check{x}^2}{s+\nu\check{x}}$ for t_2.

On the side, we have given a graphical illustration of how the parameters are updated. We used a number of pseudocounts $s\coloneqq 2$ and an observed sample $\check{x}$ of length 2; thus, taking into account the definition of τ, the two updated mean single-sample sufficient statistic parameters are

$$\begin{aligned}
t'&\coloneqq\tfrac{2}{3}\cdot t+\tfrac{1}{3}\cdot(\check{x}_1,\check{x}_1^2),\\
t''&\coloneqq\tfrac{3}{4}\cdot t'+\tfrac{1}{4}\cdot(\check{x}_2,\check{x}_2^2)=\tfrac{1}{2}\cdot t+\tfrac{1}{2}\cdot(\check{x}_1+\check{x}_2,\check{x}_1^2+\check{x}_2^2).
\end{aligned}$$



Note how the proportions appearing in these expressions also appear in the illustration.

Using (4.26)$_{167}$, (4.30), and (4.11)$_{160}$, we can obtain the prior immediate

predictive linear prevision. Its probability density for z in $\mathbb{R}$ is

$$\frac{\frac{1}{\sqrt{2\cdot\pi}}\cdot 2\cdot\frac{\left(\frac{s}{2}\cdot(t_2-t_1^2)\right)^{\frac{s+3}{2}}}{\Gamma\frac{s+3}{2}}\cdot\sqrt{\frac{s}{2\cdot\pi}}}{2\cdot\frac{\left(\frac{s+1}{2}\cdot\left(\frac{s\cdot t_2+z^2}{s+1}-(\frac{s\cdot t_1+z}{s+1})^2\right)\right)^{\frac{s+1+3}{2}}}{\Gamma\frac{s+1+3}{2}}\cdot\sqrt{\frac{s+1}{2\cdot\pi}}}$$

$$=\frac{1}{\sqrt{2\cdot\pi\cdot\frac{s+1}{s}}}\cdot\frac{\Gamma\frac{s+3+1}{2}}{\Gamma\frac{s+3}{2}}\cdot\frac{\left(\frac{s}{2}\cdot(t_2-t_1^2)\right)^{\frac{s+3+1}{2}}\cdot\left(\frac{s}{2}\cdot(t_2-t_1^2)\right)^{-\frac{1}{2}}}{\left(\frac{s+1}{2}\cdot\left(\frac{s\cdot t_2+z^2}{s+1}-(\frac{s\cdot t_1+z}{s+1})^2\right)\right)^{\frac{s+3+1}{2}}}$$

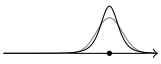
after some tedious manipulations this can be rewritten as

$$=\frac{1}{\sqrt{(s+1)\cdot(t_2-t_1^2)}}\cdot\frac{\Gamma\frac{s+3+1}{2}}{\Gamma\frac{s+3}{2}\cdot\sqrt{\pi}}\cdot\left(1+\frac{1}{(s+1)\cdot(t_2-t_1^2)}\cdot(z-t_1)^2\right)^{-\frac{s+3+1}{2}},$$

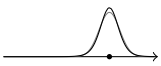
Three Student densities (plot restricted to $[-5,5]$): (mean μ indicated with a dot)

$(\alpha,\mu,\sigma)=(4,0,2)$

$(\alpha,\mu,\sigma)=(4,2,0.6)$



$(\alpha,\mu,\sigma)=(9,2,0.6)$



In grey, normal densities with the same mean and variance are given.

in which some will recognize the probability density of Student's distribution. The Student linear prevision [see, e.g., Bernardo & Smith 1994, §3.2.2$_{122}$] is defined by (let f be a measurable gamble on $\mathbb{R}$)

$$\mathrm{St}(f|\alpha,\mu,\sigma):=\int_{\mathbb{R}}fz\cdot\frac{1}{\sigma\cdot\sqrt{\alpha}}\cdot\frac{\Gamma\frac{\alpha+1}{2}}{\Gamma\frac{\alpha}{2}\cdot\sqrt{\pi}}\cdot\left(1+\frac{1}{\alpha}\cdot\left(\frac{z-\mu}{\sigma}\right)^2\right)^{-\frac{\alpha+1}{2}}\mathrm{d}z,\qquad(4.34)$$

where μ – which is the distribution's mean – must be a real number and α and σ – which is *not* the distribution's standard deviation, $\sigma\cdot\sqrt{\frac{\alpha}{\alpha-2}}$ is – must be strictly positive. (Note that Student's distribution converges to the normal distribution as $\alpha\to+\infty$.) So the immediate prior predictive linear prevision is

$$\mathrm{St}\Big(\bullet\;\Big|\;s+3,t_1,\sqrt{t_2-t_1^2}\cdot\sqrt{\tfrac{s+1}{s+3}}\Big).\qquad(4.35)$$

The posterior is obtained by the same substitution as for the conjugate posterior.

From (4.12)$_{160}$, we know that the components of the pseudomean for the normal conjugate and predictive family can be interpreted respectively as a hypothetical mean and second-order noncentral moment. This interpretation is further supported by what we can learn from (4.29)$_{168}$: the parametric prevision of the mean and the noncentral second-order moment, or, in other words, the immediate predictive prevision of τ_1 and τ_2 are the Student distribution's mean t_1 and noncentral second moment t_2, respectively. (Be aware that the parametric prevision of the variance,

$$\mathrm{Ng}\big(\mathrm{D}_2\kappa-(\mathrm{D}_1\kappa)^2\;\big|\;t_1,s,\tfrac{s+3}{2},\tfrac{s}{2}\cdot(t_2-t_1^2)\big)=\tfrac{s}{s+1}\cdot(t_2-t_1^2),$$

is *not* the Student distribution's variance. It cannot be found using the predictive prevision.)

4.1.7 *Conjugate & predictive linear previsions for Bernoulli & negative multinomial sampling*

We now continue our example of Bernoulli sampling started in §4.1.3$_{160}$.

The information in (4.15)$_{161}$ and (4.16)$_{161}$ fully defines the canonical form (4.21)$_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}^{\mathscr{X}_{\neq o}}$ – where, recall, o is some (any) category chosen for technical elimination – and the domain of possible parameters is $\operatorname{int}\mathscr{T} = \{t : (\mathbb{R}_{>0})^{\mathscr{X}_{\neq o}} \mid \sum t < 1\}$. The one thing we do not yet know explicitly is the normalization factor (4.22)$_{165}$: (let s be a strictly positive real number and $t : \operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\Xi} \big(1 + \sum_{z:\mathscr{X}_{\neq o}} \exp \xi_z\big)^{-s} \cdot \exp^s \langle \xi | t \rangle \, \mathrm{d}\xi.$$

It can be calculated more easily when starting from the original non-canonical parameterization $\psi = \vartheta : \operatorname{int}\Delta_{\mathscr{X}} ; \big(\ln \frac{\vartheta_z}{\vartheta_o} \mid z : \mathscr{X}_{\neq o}\big)$, where ϑ is the Bernoulli distribution's frequency vector (completely determined by $\vartheta_{\mathscr{X}_{\neq o}}$):

$$\begin{aligned}
\mathsf{c}(s,t) &= 1/\textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} \vartheta_o^s \cdot \exp^s\big(\sum_{z:\mathscr{X}_{\neq o}} \ln \frac{\vartheta_z}{\vartheta_o} \cdot t_z\big) \cdot \big|\big|(\nabla\psi)\vartheta\big|\big| \, \mathrm{d}\vartheta \\
&= 1/\textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} \vartheta_o^s \cdot \Big(\prod_{z:\mathscr{X}_{\neq o}} \big(\frac{\vartheta_z}{\vartheta_o}\big)^{s\cdot t_z}\Big) \cdot \big|\big|(\mathbb{1}\vartheta_{\mathscr{X}_{\neq o}})^{-1} + \tfrac{1}{\vartheta_o}\big|\big| \, \mathrm{d}\vartheta \\
&= 1/\textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} \vartheta_o^{s\cdot(1-\sum t)} \cdot \Big(\prod_{z:\mathscr{X}_{\neq o}} \vartheta_z^{s\cdot t_z}\Big) \\
&\qquad\qquad \cdot \big|\big|(\mathbb{1}\vartheta_{\mathscr{X}_{\neq o}})^{-1}\big| \cdot \big|\mathbb{1} + \vartheta_{\mathscr{X}_{\neq o}} (\mathscr{X}_{\neq o} ; \tfrac{1}{\vartheta_o})^{\intercal}\big|\big| \, \mathrm{d}\vartheta;
\end{aligned}$$

Matrix notation: identity matrix $\mathbb{1}$, matrix product $\bullet\bullet$, transposition $\bullet^{\intercal}$, dyadic product $\bullet\bullet^{\intercal}$, and scalar product $\bullet^{\intercal}\bullet$.

writing out the first determinant and using Sylvester's determinant theorem [a particularization of Bernstein 2005, §2.13.2$_{60}$] on the second, we find

$$\begin{aligned}
&= 1/\textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} \vartheta_o^{s\cdot(1-\sum t)} \cdot \Big(\prod_{z:\mathscr{X}_{\neq o}} \vartheta_z^{s\cdot t_z}\Big) \\
&\qquad\qquad \cdot \Big(\prod_{z:\mathscr{X}_{\neq o}} \vartheta_z^{-1}\Big) \cdot |1 + (\mathscr{X}_{\neq o} ; \tfrac{1}{\vartheta_o})^{\intercal} \vartheta_{\mathscr{X}_{\neq o}}| \, \mathrm{d}\vartheta;
\end{aligned}$$

as $1 + (\mathscr{X}_{\neq o} ; \frac{1}{\vartheta_o})^{\intercal} \vartheta_{\mathscr{X}_{\neq o}} = 1 + \frac{1}{\vartheta_o} \cdot \sum_{z:\mathscr{X}_{\neq o}} \vartheta_z = 1 + \frac{1-\vartheta_o}{\vartheta_o} = \frac{1}{\vartheta_o}$, this becomes

$$\begin{aligned}
&= 1/\textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} \vartheta_o^{s\cdot(1-\sum t)-1} \cdot \Big(\prod_{z:\mathscr{X}_{\neq o}} \vartheta_z^{s\cdot t_z - 1}\Big) \, \mathrm{d}\vartheta \\
&= \frac{\Gamma s}{\Gamma(s - \sum s\cdot t) \cdot \prod_{z:\mathscr{X}_{\neq o}} \Gamma(s\cdot t_z)}.
\end{aligned} \tag{4.36}$$

The last step follows from recognizing the Dirichlet integral (which we already encountered in (3.110)$_{139}$). However, instead of being defined for all continuous gambles on the *closed* unit simplex $\Delta_{\mathscr{X}}$, it is now defined for all measurable gambles f on the *open* unit simplex $\operatorname{int}\Delta_{\mathscr{X}}$:

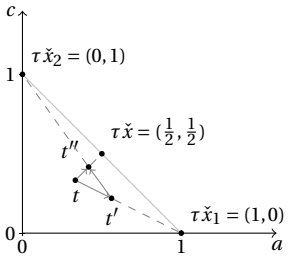
$$\mathrm{Dj}(f|\alpha) \coloneqq \tfrac{\Gamma(\sum\alpha)}{\prod_{z:\mathscr{X}}\Gamma\alpha_z} \cdot \textstyle\int_{\operatorname{int}\Delta_{\mathscr{X}}} f\vartheta \cdot \Big(\prod_{z:\mathscr{X}} \vartheta_z^{\alpha_z - 1}\Big) \, \mathrm{d}\vartheta, \tag{4.37}$$

where α is a strictly positive real vector on $\mathscr{X}$.

So, as we already knew, the conjugate family is the Dirichlet family. The technical difference in domain between the one obtained here and the one found in §3.3.1$_{139}$ is due to the approach taken. The prior parametric linear prevision is

$$\mathrm{Dj}\big(\bullet \;\big|\; (s\cdot t, s-\textstyle\sum s\cdot t)\big), \tag{4.38}$$

where it should be understood that the parameter's last component is the o-component. The posterior is obtained by substituting $s+\nu\check{x}$ for s and $s\cdot t+(C_{\mathscr{X}}\check{x})_{\mathscr{X}_{\neq o}}$ for $s\cdot t$.



On the side, we have given a graphical illustration of how the parameters are updated. We used a number of pseudo-counts $s:=2$ and an observed sample $\check{x}:=\check{x}_1\check{x}_2=ac$; thus, taking into account the definition of τ, the two updated mean single-sample sufficient statistic parameters are

$$\begin{aligned} t' &:= \tfrac{2}{3}\cdot t+\tfrac{1}{3}\cdot(1,0),\\ t'' &:= \tfrac{3}{4}\cdot t'+\tfrac{1}{4}\cdot(0,1)=\tfrac{1}{2}\cdot t+\tfrac{1}{2}\cdot(\tfrac{1}{2},\tfrac{1}{2}). \end{aligned}$$

Note how the proportions appearing in these expressions also appear in the illustration.

Using (4.27)$_{167}$, (4.36)$_{\curvearrowright}$, and (4.16)$_{161}$, we can obtain the prior immediate predictive prevision. Its probability mass for z in $\mathscr{X}_{\neq o}$ is

$$\frac{1\cdot\dfrac{\Gamma s}{\Gamma(s-\sum s\cdot t)\cdot\prod_{z':\mathscr{X}_{\neq o}}\Gamma(s\cdot t_{z'})}}{\dfrac{\Gamma(s+1)}{\Gamma(s-\sum s\cdot t)\cdot\prod_{z':\mathscr{X}_{\neq o}}\Gamma\big(s\cdot t_{z'}+(C_{\mathscr{X}}z)_{\mathscr{X}_{\neq o}}\big)}} = \tfrac{\Gamma s}{\Gamma(s+1)}\cdot\frac{\prod_{z':\mathscr{X}_{\neq o}}\Gamma(s\cdot t_{z'}+\delta_{zz'})}{\prod_{z':\mathscr{X}_{\neq o}}\Gamma(s\cdot t_{z'})};$$

and thus, because $\Gamma(\alpha+1)=\alpha\cdot\Gamma\alpha$ for any real α,

$$=\tfrac{1}{s}\cdot s\cdot t_z=t_z.$$

Similarly, the probability mass for o is $1-\sum t$. so the prior immediate predictive prevision is the weighted average prevision (cf. (3.90)$_{128}$)

$$\mathrm{Wa}\big(\bullet \;\big|\; (t, 1-\textstyle\sum t)\big). \tag{4.39}$$

The posterior is obtained by the same substitution as for the conjugate posterior.

From (4.17)$_{162}$, we know that the pseudomean for the Bernoulli conjugate and predictive family can be interpreted as a hypothetical frequency vector. This interpretation is further supported by what we can learn from (4.29)$_{168}$: the parametric prevision of this frequency vector, or, in other words, the immediate predictive prevision of τ, is the weighted average's (normalized) vector of weights t.

Not only the previsions for immediate prediction can be derived using $(4.27)_{167}$, of course. But immediate predictive previsions can also be derived in terms of a noncanonical parameterization, mirroring $(4.25)_{166}$. For example, given some count vector m in $\mathbb{N}^{\mathscr{X}}$, we can use

$$\mathrm{Dj}\big(\mathrm{Cn}(\bullet|m,\bullet)\,\big|\,(s\cdot t, s-\textstyle\sum s\cdot t)\big)$$

as a predictive linear prevision for negative count-multinomial sampling (cf. $(4.19)_{163}$). Also see $\S3.3.1_{139}$ to spot the parallels with how the Dirichlet-multinomial as the predictive prevision for multinomial sampling can be seen as a combination of the Dirichlet prevision and multinomial prevision.

4.2 Imprecise-probabilistic inference models for exponential families

Now that we have had an overview of the theory of exponential families and the related conjugate parametric and predictive families, we are ready to use those related families to construct imprecise-probabilistic parametric and predictive inference models for exponential families. This is done in the first subsection; in the two other subsections our two examples continue.

4.2.1 *Parametric & predictive inference models*

We consider a process that is assumed to generate samples that are distributed according to an unknown member of a well-specified regular exponential family distribution (cf. $\S4.1.1_{155}$), so the functions a, τ, and ψ defining the exponential family are known, but the parameter value ϕ defining the actual member is not. In $\S4.1.4_{164}$ and $\S4.1.5_{166}$, we have made the acquaintance of its conjugate parametric and predictive families. The members of these families come in pairs, consisting of a linear prevision from each defined by the same hyperparameters: a number of counts – a strictly positive real number, interpretable as a sample size – and some scalar or vector that can be interpreted as a mean single-sample sufficient statistic.

The parametric prevision can be used to describe our subject's uncertainty about the parameter value and the predictive prevision to describe his uncertainty about the value of future samples. Before observing any sample, prior previsions $\mathrm{Cf}^{\mathsf{a},\tau}(\bullet|s,t)$ or $\mathrm{Cf}^{\mathsf{a},\tau,\psi}(\bullet|s,t)$ and $\mathrm{Pf}^{\mathsf{a},\tau}(\bullet|\mathscr{Y},s,t)$, where $\mathscr{Y} :\subset \mathscr{X}^*$, or $\mathrm{Pf}^{\mathsf{a},\tau}(\bullet|s,t)$ would be used, characterized by a number of pseudocounts s and a pseudomean t. After the observation of a sample sequence $\check{x}$, these priors can be updated to posteriors with the same functional form, but with respective updated pseudocounts $s+v\check{x}$ and pseudomean $\frac{s\cdot t+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}$.

At this point, the question becomes: how should our subject choose his number of pseudocounts s and his pseudomean t? Actually, considering $\mathrm{Pf}^{\mathsf{a},\tau}(\tau|s,t) = t$ by $(4.29)_{168}$, it would only be reasonable to be so specific as to restrict his choice to only a single t in $\operatorname{int}\mathscr{T}$ if he is prepared to bet accordingly precise at any for him reasonable stake. This should only be the case when he has sufficient information supporting this precise assessment. Quite often this is not the case, and his initial assessment is better captured by bounds, i.e., supremum acceptable buying prices and infimum acceptable selling prices for the component gambles of τ. This would result in an initial assessment represented by a convex subset of the set $\operatorname{int}\mathscr{T}$ of mean single-sample sufficient statistics. Any other (possibly nonconvex) subset of $\operatorname{int}\mathscr{T}$ could also be used, as a result of some convenient ad-hoc reasoning or of a more complicated elicitation procedure involving assessments other than just the predictive assessments for the components of τ.

The number of pseudocounts could be seen as the weight accorded to the initial assessment. Considering that for a given pseudomean the posterior prevision is

$$\mathrm{Pf}^{\mathsf{a},\tau}\left(\tau \,\middle|\, s+v\check{x}, \tfrac{s\cdot t+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}\right) = \tfrac{s\cdot t+v\check{x}\cdot\tau\check{x}}{s+v\check{x}} = \tfrac{s}{s+v\check{x}}\cdot t + \tfrac{v\check{x}}{s+v\check{x}}\cdot\tau\check{x},$$

the chosen number s determines what could be seen as the speed of learning, i.e., how many samples need to be observed before the influence of the pseudomean in the prevision is equal to the influence of the observed mean single-sample sufficient statistic. Based on this, our subject must make a choice for s that is appropriate for the application at hand; to me it seems that in general any more specific advice cannot be given.

Now the moment has come to introduce the promised imprecise-probabilistic parametric and predictive inference models – or, respectively, ICEFM and IPEFM , for short. As explained, a number of pseudocounts $s:\mathbb{R}_{>0}$ must be chosen as well as a subset $\mathscr{U}$ of $\operatorname{int}\mathscr{T}$. This determines the prior parametric and predictive lower previsions using lower envelopes: (let $\mathscr{Y} :\subset \mathscr{X}^*$)

$$\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(\bullet|s,\mathscr{U}) := \inf_{t:\mathscr{U}} \mathrm{Cf}^{\mathsf{a},\tau}(\bullet|s,t), \tag{4.40}$$

$$\underline{\mathrm{Cf}}^{\mathsf{a},\tau,\psi}(\bullet|s,\mathscr{U}) := \inf_{t:\mathscr{U}} \mathrm{Cf}^{\mathsf{a},\tau,\psi}(\bullet|s,t), \tag{4.41}$$

$$\underline{\mathrm{Pf}}^{\mathsf{a},\tau}(\bullet|\mathscr{Y},s,\mathscr{U}) := \inf_{t:\mathscr{U}} \mathrm{Pf}^{\mathsf{a},\tau}(\bullet|\mathscr{Y},s,t), \tag{4.42}$$

$$\underline{\mathrm{Pf}}^{\mathsf{a},\tau}(\bullet|s,\mathscr{U}) := \inf_{t:\mathscr{U}} \mathrm{Pf}^{\mathsf{a},\tau}(\bullet|s,t). \tag{4.43}$$

The posterior lower previsions after observing some sequence $\check{x}$ in $\mathscr{X}^*$ are obtained by substituting $s+v\check{x}$ for s and $\frac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}$ for $\mathscr{U}$, where we have used elementwise addition. As lower envelopes of standard Bayesian models, all these lower previsions are coherent [Walley 1991, §7.8.1_{397}].

In the literature, relatively little can be found on inference models with imprecision for sampling from exponential families (but of course much more can be found about specific families). For comparative purposes, let us quickly mention two ideas:

Walley [1991, §4.6.8$_{205}$] also formulated a number of ideas about using sets of conjugate priors.

(i) Boratyńska [1997] – working in a robust Bayesian context – considers a conjugate linear prior by fixing $s \cdot t$ and then performs a sensitivity analysis by varying this quantity.

(ii) Coolen [1993] – working in an imprecise probabilities context – uses proportional lower and upper conjugate density functions to define an imprecise-probabilistic inference model.

Both papers consider one-parameter families.

There are a number of remarks we can make about s, $\mathscr{U}$ and their influence on the model:

(i) Independent of whether $\mathscr{U}$ is convex or not, the prior and posterior parametric and predictive credal sets are of course closed convex sets. Even though $\frac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}$ may be a subset of $\mathscr{U}$, the posterior credal sets will in general be disjunct from the prior ones through the effect of the modification of the count parameter to $s+v\check{x}$.

(ii) It is important to realize that $\frac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}$, the posterior set of mean single-sample sufficient statistics, will remain unbounded in the directions of $\mathbb{R}^d$ for which $\mathscr{U}$ is. A consequence is that there will be a class of gambles for which no learning will take place, i.e., the posterior lower or upper prevision (or both) of these gambles will not differ from the prior ones. The prime example of this is the predictive lower and upper prevision of the components of τ corresponding to the unbounded directions: Let $i : 1..d$ be the index of a direction that is, e.g., unbounded from above, then symbolically

$$\begin{aligned}\overline{\mathrm{Pf}}^{\mathfrak{a},\tau}\left(\tau_i \,\middle|\, s+v\check{x}, \tfrac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}\right) &= \sup_{t:\mathscr{U}} \tfrac{s\cdot t_i+v\check{x}\cdot\tau_i\check{x}}{s+v\check{x}} \\ &= \sup_{t:\mathscr{U}} \tfrac{s+v\check{x}\cdot\tau_i\check{x}/t_i}{s+v\check{x}}\cdot t_i \\ &= \sup_{t:\mathscr{U}} \tfrac{s}{s+v\check{x}}\cdot t_i \\ &= \sup_{t:\mathscr{U}} t_i = \overline{\mathrm{Pf}}^{\mathfrak{a},\tau}(\tau_i|s,\mathscr{U}).\end{aligned}$$

(iii) Quite often, the initial assumptions made and the information available correspond to a state of so-called prior ignorance in which taking $\mathscr{U}$ equal to $\operatorname{int}\mathscr{T}$ would be the reasonable thing to do; the corresponding prior is then called a near-ignorance model [for a more general definition, see Walley 1991, §4.6.9$_{206}$,§5.3.2$_{218}$]. But because no learning takes place for unbounded directions in $\mathscr{U}$, such near-ignorance models are only really practical whenever $\mathscr{T}$ is bounded. So for exponential families for which $\mathscr{T}$ has unbounded directions, an effort should be made to arrive at some finite bound, however large, if one wishes learning to take place in the unbounded directions of $\mathscr{T}$.

More than ten years ago, Luis Raúl Pericchi already reflected on this issue [Walley 1996, p. 48].

(iv) We have already seen that the number of pseudocounts s characterizes the speed of learning, i.e., which amount of observations carries the same weight as the initial assessments. The number of pseudocounts also determines the evolution of the imprecision of the inferences generated by the model. To be precise, the relative volume $\frac{s}{s+v\check{x}}$ of the posterior set of mean single-sample sufficient statistics to the prior one can be seen as a stakes-independent measure of the imprecision. For example, (let $i:1..d$)

$$\begin{aligned}[\underline{\mathrm{Pf}}^{\mathsf{a},\tau}\big(\tau_i \,\big|\, s+v\check{x}, \tfrac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}\big), \overline{\mathrm{Pf}}^{\mathsf{a},\tau}\big(\tau_i \,\big|\, s+v\check{x}, \tfrac{s\cdot\mathscr{U}+v\check{x}\cdot\tau\check{x}}{s+v\check{x}}\big)]\\ = \tfrac{s}{s+v\check{x}}\cdot[\inf_{t:\mathscr{U}} t_i, \sup_{t:\mathscr{U}} t_i] + \tfrac{v\check{x}}{s+v\check{x}}\cdot\tau_i\check{x}\\ = \tfrac{s}{s+v\check{x}}\cdot[\underline{\mathrm{Pf}}^{\mathsf{a},\tau}(\tau_i|s,\mathscr{U}), \overline{\mathrm{Pf}}^{\mathsf{a},\tau}(\tau_i|s,\mathscr{U})] + \tfrac{v\check{x}}{s+v\check{x}}\cdot\tau_i\check{x}.\end{aligned}$$

This example must not, however, leave the impression that this imprecision measure determines the imprecision for all gambles (of both the parametric and predictive models). There are even situations (exponential families and generated sample sequences) and gambles for which the imprecision increases; an example is given in the Bestiarium, near the end of §B.1.4$_{210}$. It is still unclear if the proposed inference models are dilation prone [Seidenfeld & Wasserman 1993], i.e., if there are gambles for which the imprecision increases no matter what single-sample observation has been made.

Illustrating the closure operator: $\mathrm{cl}]0,1[= [0,1]$.

(v) Whenever $\tau\check{x} \notin \mathrm{cl}\,\mathscr{U}$ when starting from a state of prior ignorance, there is so-called prior-data conflict [see, e.g., Walley 1991, §1.1.4$_6$], the more so the further $\tau\check{x}$ is from $\mathscr{U}$. Apart from getting more samples and hoping that the conflict will disappear, there are two immediate options for modifying the inference model to deal with the conflict (by increasing the imprecision):
   (a) Enlarging $\mathscr{U}$ to encompass $\tau\check{x}$; this seems appropriate when, after observing the sample, the subject realizes his bounds were not chosen wide enough.
   (b) Allowing the number of pseudocounts to vary (e.g., in an interval $]0,u]$, with $u:\mathbb{R}_{>0}$), which allows the weight accorded to the conflict-creating hypothetical initial sample to be reduced [Walley 1991, §5.4$_{222}$].

   Currently, I prefer the first option, mainly because the possible effects of the second one are not yet clear to me; research into this second option is currently being done by Walter & Augustin [2009].

(vi) Why not vary (s,t) over a subset of $\mathbb{R}_{>0}\times\operatorname{int}\mathscr{T}$ instead of just varying t in $\mathscr{T}$? We were led to this choice because of the intuitive interpretation attached to both parameters: s as a hypothetical number of counts determining the learning speed – of which a unique choice for the whole model seems natural –, t as a mean

single-sample sufficient statistic – which through the posterior weighted average prediction property $(4.29)_{168}$ allows for a straightforward expression of prior uncertainty. Of course it is also mathematically convenient to not consider subsets of $\mathbb{R}_{>0} \times \operatorname{int}\mathscr{T}$: we avoid ending up with more complex models.

We have said that the expressions (4.40)–$(4.43)_{174}$ for the prior lower previsions of the ICEFM and IPEFM are but a substitution away from the posterior lower previsions. However, intuitively appealing though it may be, we have not yet justified this. What we need to show is that this follows from one of the two updating procedures, natural extension $(1.83)_{61}$ or regular extension $(1.70)_{58}$. Just below $(4.24)_{166}$ we mentioned that $\mathrm{Cf}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,t) > 0$ for any t in $\operatorname{int}\mathscr{T}$. So in any case $\overline{\mathrm{Cf}}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,\mathscr{U}) > 0$. When $\operatorname{cl}\mathscr{U} \subseteq \operatorname{int}\mathscr{T}$, then $\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,\mathscr{U}) > 0$, so in this case natural extension would result in the substitution posterior, as then

$$\operatorname{ext}\bigl(\mathscr{M}\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(\bullet|s,\mathscr{U})\bigr) \subseteq \bigl\{\mathrm{Cf}^{\mathsf{a},\tau}(\bullet|s,t) \bigm| t : \operatorname{cl}\mathscr{U}\bigr\}.$$

However, when $\operatorname{cl}\mathscr{U} \cap (\mathscr{T} \setminus \operatorname{int}\mathscr{T}) \neq \emptyset$, then possibly $\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,\mathscr{U}) = 0$, in which case natural extension would not give the substitution posterior, but the vacuous lower prevision. Luckily for us, regular extension always does what we want: As

$$\bigl\{\mathrm{Cf}^{\mathsf{a},\tau}(\bullet|s,t) \bigm| t : \mathscr{U}\bigr\} \subset \bigl\{R : \mathscr{M}\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(\bullet|s,\mathscr{U}) \bigm| RE^{\mathsf{a},\tau}_{\check{x}} > 0\bigr\},$$

we know that the substitution posterior dominates the regular extension. But we can say something even stronger: they coincide. To wit, the substitution posterior is coherent and for any measurable gamble f on Ξ it holds that $\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(f|s+\nu\check{x}, \frac{s\cdot\mathscr{U}+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}) \geq 0$ when $\mathrm{Cf}^{\mathsf{a},\tau}(E^{\mathsf{a},\tau}_{\check{x}}|s,t) > 0$ and $\underline{\mathrm{Cf}}^{\mathsf{a},\tau}(f|s,\mathscr{U}) \geq 0$. These two properties *characterize* the regular extension [Walley 1991, end of §J3_{640}].

We close this subsection by generalizing the posterior weighted average prediction property $(4.29)_{168}$ to our imprecise-probabilistic inference models. Let $i : 1..d$, then we have

$$\underline{\mathrm{Pf}}^{\mathsf{a},\tau}(\tau_i|s,\mathscr{U}) = \underline{\mathrm{Cf}}^{\mathsf{a},\tau}(\mathrm{D}_i\kappa|s,\mathscr{U}) = \inf_{t:\mathscr{U}} t_i, \tag{4.44}$$

$$\overline{\mathrm{Pf}}^{\mathsf{a},\tau}(\tau_i|s,\mathscr{U}) = \overline{\mathrm{Cf}}^{\mathsf{a},\tau}(\mathrm{D}_i\kappa|s,\mathscr{U}) = \sup_{t:\mathscr{U}} t_i. \tag{4.45}$$

We call this the posterior contamination prediction property: the posterior predictive lower prevision of a single-sample sufficient statistic's i-th component is $\frac{\nu\check{x}}{s+\nu\check{x}} \cdot \tau_i\check{x} + \frac{s}{s+\nu\check{x}} \cdot \inf_{t:\mathscr{U}} t_i$, i.e., the contamination of the observed mean single-sample sufficient statistic $\tau\check{x}$ with $[\inf_{t:\mathscr{U}} t_i, \sup_{t:\mathscr{U}} t_i]$.

For gambles that can be written as affine combinations of the components of τ – which, recall remark $(\mathrm{iii})_{157}$, are affinely independent – or, equivalently, of the components of $\nabla\kappa$, this property allows for a greatly increased efficiency in the computation of their prevision. Let α be a

vector of coefficients in $\mathbb{R}^d$ and β a real number, then

$$\begin{aligned}\underline{\mathrm{Pf}}^{\mathsf{a},\tau}\big(\beta+\textstyle\sum\alpha\cdot\tau \,\big|\, s+\nu\check{x}, \frac{s\cdot\mathscr{U}+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}\big)&\\ =\underline{\mathrm{Cf}}^{\mathsf{a},\tau}\big(\beta+\textstyle\sum\alpha\cdot\nabla\kappa \,\big|\, s+\nu\check{x}, \frac{s\cdot\mathscr{U}+\nu\check{x}\cdot\tau\check{x}}{s+\nu\check{x}}\big)&\\ =\beta+\tfrac{\nu\check{x}}{s+\nu\check{x}}\cdot\textstyle\sum\alpha\cdot\tau\check{x}+\tfrac{s}{s+\nu\check{x}}\cdot\inf_{t:\mathscr{U}}\sum\alpha\cdot t.&\end{aligned} \tag{4.46}$$

We see that the only computationally nontrivial part of the right-hand side expression is $\inf_{t:\mathscr{U}}\sum\alpha\cdot t$, i.e., to optimize a linear function over $\mathscr{U}$. Knowing that $\mathscr{T}$ itself is convex (cf. §4.1.1$_{155}$) and having some leeway in choosing the bounds that define $\mathscr{U}$, it is a good move to let these bounds be defined by convex functions. This guarantees that the above optimization problem is a convex optimization problem, for which many computationally efficient algorithms exist [see, e.g., Boyd & Vandenberghe 2004]. Of course, this may also be the case for other types of gambles than these affine combinations, but these are not as easily identifiable.

There is an interesting balance with regard to the optimization problems that need to be solved when calculating immediate predictive lower (and upper) previsions (4.43)$_{174}$ of our inference models: The more complex a model is – i.e., the higher the number $d+1$ of scalar parameters relative to the size (cardinality or dimension) of the sample space $\mathscr{X}$ –, the larger the space $\mathrm{span}\big(\{\tau\}\cup\iota(\mathscr{X};1)\big)$ of gambles is for which these optimization problems are efficiently solvable. For parametric lower previsions (4.40)$_{174}$ the dimension d of the domain Ξ of the objective functions always is practically the same as the dimension $d+1$ of $\mathrm{span}\big(\{\nabla\kappa\}\cup\iota(\Xi;1)\big)$.

This concludes the theoretical presentation of the ICEFM and IPEFM. What remains to be done is first illustrate what we have seen here in the next two subsections and then, in the next section illustrate how these models can be used in an application.

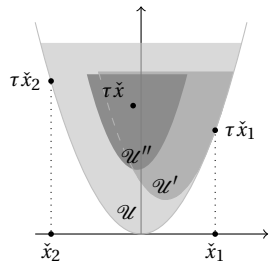
4.2.2 *Inference models for normal sampling*

We again take up our normal sampling example previously looked at in §4.1.2$_{159}$ and §4.1.6$_{168}$. We already know all the functions and sets ((4.10)$_{160}$ and (4.11)$_{160}$) that characterize the normal family as a regular exponential family and we have discovered that the conjugate linear prevision (4.33)$_{169}$ is a normal-gamma prevision and that the immediate predictive linear prevision (4.35)$_{170}$ is a Student prevision.

Then all it takes to define imprecise-probabilistic inference models of the type of (4.40)–(4.43)$_{174}$ for normal sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T}=\{t:\mathbb{R}^2 \mid t_2>t_1^2\}$. The choice of bounds can be guided by the interpretation given to the components of $t:\mathscr{T}$. Bounding t_1 corresponds to bounding the mean; bounding t_2 corresponds to bounding the second-order noncentral moment; bounding $t_2-t_1^2$ corresponds to bounding the variance (times $\frac{s+1}{s}$,

cf. just above §4.1.7$_{171}$). Combining the last and the first, which might be an intuitively attractive idea, gives rise to a nonconvex set $\mathscr{U}$ and is therefore less than ideal from an optimization viewpoint.

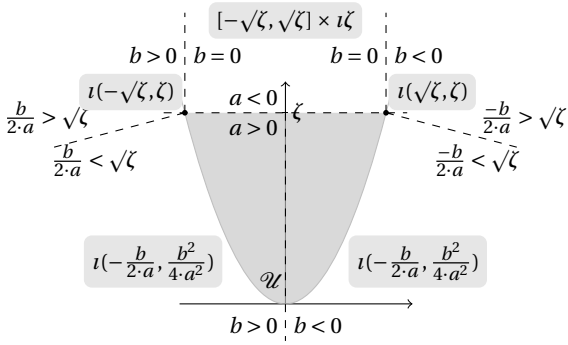
On the side, we again give a graphical illustration of how the parameters are updated (cfr. the corresponding illustration in §4.1.6$_{168}$). We use a number of pseudocounts $s := 2$, the initial set of mean single-sample sufficient statistics $\mathscr{U}$ is defined by choosing an upper bound for the second component (i.e., bounding the second-order noncentral moment), and we take an observed sample $\check{x}$ of length 2. Thus, taking into account the definition of τ, the two updated sets of mean single-sample sufficient statistics are



$$\mathscr{U}' := \tfrac{2}{3}\cdot\mathscr{U} + \tfrac{1}{3}\cdot(\check{x}_1, \check{x}_1^2),$$
$$\mathscr{U}'' := \tfrac{3}{4}\cdot\mathscr{U}' + \tfrac{1}{4}\cdot(\check{x}_2, \check{x}_2^2) = \tfrac{1}{2}\cdot\mathscr{U} + \tfrac{1}{2}\cdot(\check{x}_1 + \check{x}_2, \check{x}_1^2 + \check{x}_2^2).$$

Note again how the proportions appearing in these expressions are reflected in the illustration.

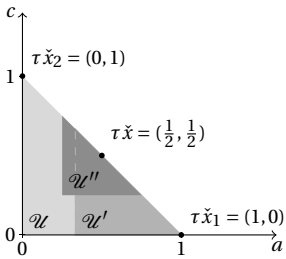
Due to the fact that $\tau = z : \mathbb{R} ; (z, z^2)$, the computation of the immediate predictive lower and upper previsions of all quadratic gambles on $\mathbb{R}$ can be efficiently handled whenever $\mathscr{U}$ is convex. For example, when taking $\mathscr{U}$ as in the illustration above, with an upper bound $\zeta : \mathbb{R}_{>0}$, i.e., equal to $\{t : \mathbb{R}^2 \mid \zeta > t_2 > t_1^2\}$, and looking at a gamble $a\cdot\tau_2 + b\cdot\tau_1 + c$, where a, b, and c are real numbers, the optimization problem to solve for finding the immediate prevision of this (continuous) gamble is essentially $\min_{t:\mathrm{cl}\,\mathscr{U}} a\cdot t_2 + b\cdot t_1$ (cf. (4.46)). The straightforward solution to this problem is summarized in the illustration next to this: $\mathscr{U}$ can be partitioned in a number of subsets (using dashed lines); each of these contains the solution set to a subset of the possible (a, b)-values (presented on a light gray background).



4.2.3 *Inference models for Bernoulli sampling*

We now take up our Bernoulli sampling example, previously looked at in §4.1.3$_{160}$ and §4.1.7$_{171}$. We already know all the functions and sets ((4.15)$_{161}$ and (4.16)$_{161}$) that characterize the Bernoulli family as a regular exponential family and we have rediscovered that the conjugate linear prevision (4.38)$_{172}$ is a Dirichlet prevision and that the immediate predictive linear prevision (4.39)$_{172}$ is a weighted average prevision.

Then again all it takes to define imprecise-probabilistic inference models of the type of (4.40)–(4.43)$_{174}$ for Bernoulli sampling is to choose a bounded subset of $\operatorname{int}\mathscr{T} = \left\{t : (\mathbb{R}_{>0})^{\mathscr{X}_{\neq o}} \,\middle|\, \sum t < 1\right\}$ and a number of pseudocounts s in $\mathbb{R}_{>0}$. As $\operatorname{int}\mathscr{T}$ already is a bounded set, the easiest choice is take $\mathscr{U} := \operatorname{int}\mathscr{T}$ and start with a near-ignorance prior, unless there is so much prior information about the Bernoulli process that would make it worthwhile to specify bounds.



On the side, we again give a graphical illustration of how the parameters are updated (cfr. the corresponding illustration in §4.1.7$_{171}$). We use a number of pseudocounts $s := 2$, the initial set of mean single-sample sufficient statistics is $\mathscr{U} := \operatorname{int}\mathscr{T}$, which corresponds to a near-ignorance prior, and we take an observed sample $\check{x} := \check{x}_1\check{x}_2 = ac$. Thus, taking into account the definition of τ, the two updated sets of mean single-sample sufficient statistics are

$$\mathscr{U}' := \tfrac{2}{3}\cdot\mathscr{U} + \tfrac{1}{3}\cdot(1,0),$$
$$\mathscr{U}'' := \tfrac{3}{4}\cdot\mathscr{U}' + \tfrac{1}{4}\cdot(0,1) = \tfrac{1}{2}\cdot\mathscr{U} + \tfrac{1}{2}\cdot(\tfrac{1}{2},\tfrac{1}{2}).$$

Note again how the proportions appearing in these expressions are reflected in the illustration.

As the components of $\tau = z : \mathscr{X} ; (C_{\mathscr{X}} z)_{\mathscr{X}_{\neq o}}$ form the set $\{I^{z'} \mid z' : \mathscr{X}_{\neq o}\}$ of indicator functions on $\mathscr{X}$, their affine hull is the set of all gambles on $\mathscr{X}$. So then we know from (4.46)$_{178}$ that the computation of the immediate predictive lower and upper previsions of all gambles can be handled efficiently. This is not a big surprise: we already know that this prevision is a linear-vacuous mixture.

4.3 APPLICATION: NAIVE CREDAL CLASSIFICATION

In this section, as an illustration of a practical application of the IPEFM introduced in the previous section, we explain how they can be used for credal classification. In the first two subsections, §4.3.1 and §4.3.2$_{182}$, we describe what we understand by classification and zoom in, first on credal classifiers and then on naive credal classifiers. This last type of classifier, as introduced by Zaffalon [2001], is what we are going to generalize in §4.3.3$_{186}$, the last subsection, by replacing some of the predictive inference models that form its building blocks by IPEFMs.

Of course, apart from using these inference models as such and for credal classification, other potential applications exist. We should at least mention one: the application to linear regression by Walter et al. [2007].

4.3.1 *Credal classification*

Let us first begin by explaining what we understand by classification: This is the task of assigning one or more *predefined* classes to a – possibly

nonphysical – object. We assume that this object is represented by a tuple of attributes, which are labels or numerical values that each characterize some aspect of the object. So, formally, for us a classifier is a function that maps attribute tuples belonging to some attribute set $\mathfrak{A}$ to a subset of the discrete set $\mathfrak{C}$ of predefined classes.

Michie et al. [1994] provide a nice overview of classification procedures.

Of course, not just any function in $\mathfrak{A} \to \wp\mathfrak{C}$ is a good classifier. Only those classifiers are considered to be good that manage to assign sets of classes to an object that are considered fitting by some external criterion. This external criterion can be classification by a human or a later revelation of the true class through the discovery or appearance of some extra attribute(s). We are not going to discuss these external criteria here.

An example classification task could consist of finding the disease that corresponds with a number of symptoms observed in a patient. The true disease is ideally in the set of diseases picked by the classifier; it could be revealed by a positive response to some treatment or a post-mortem examination.

The type of classifier we are going to look at – the credal classifier – uses, to classify an object with attribute tuple $\mathfrak{a}$ in $\mathfrak{A}$, a coherent lower classifying prevision $\underline{E}(\bullet|\mathfrak{a})$ on $\mathscr{L}_{\mathfrak{C}}$; it expresses our uncertainty about the class the object belongs to. To classify the object, we use this prevision and some decision criterion. A possible, simple criterion is interval dominance [see, e.g., Troffaes 2007]: choose the set consisting of those classes for which the upper probability is never lower than the lower probability of any of the other classes.

The criterion we are going to use, however, is maximality (for a description, see §3.4.2$_{144}$, paragraphs three and four). For this, we need to associate a utility function (also known as a gain function, or, via its negation, a loss function) with every class $\mathfrak{c}$ in $\mathfrak{C}$; it is a gamble $U_{\mathfrak{c}}$ on $\mathfrak{C}$ that returns the utility $U_{\mathfrak{c}}\mathfrak{d}$ when the object is classified in the class $\mathfrak{c}$ and turns out to belong to the class $\mathfrak{d}$. Once these utility functions have been chosen, the credal classifier generates the set of those classes $\mathfrak{c}$ for which $\min_{\mathfrak{d}:\mathfrak{C}} \overline{E}(U_{\mathfrak{c}} - U_{\mathfrak{d}}|\mathfrak{a}) \geq 0$. In other words: a pairwise comparison between all classes $\mathfrak{d}$ and $\mathfrak{e}$ is performed by seeing how $\underline{E}(U_{\mathfrak{d}} - U_{\mathfrak{e}}|\mathfrak{a})$ strictly compares to 0 and then the classifier generates the maximal elements of the resulting strict partial order.

When for each attribute tuple $\mathfrak{a}$ a linear prevision is used instead of the coherent lower prevision $\underline{E}(\bullet|\mathfrak{a})$ on $\mathscr{L}_{\mathfrak{C}}$, the classifier can be called a Bayesian classifier. In this case, interval dominance reduces to plain dominance – i.e., maximizing probability – and for maximality, the order becomes total and the classifier selects to classes that maximize expected utility [Troffaes 2007]. In both cases, the classifier only generates singletons or sets of classes that are deemed equivalent, in contrast to credal classifiers, which can and often do generate sets of classes that are deemed incomparable. This is also the main advantage of credal

classifiers over Bayesian classifiers: they can take the *amount of data* in the learning set into account and give cautious classifications (sets containing a relatively large number of classes) when relatively little data is available.

Until now, we have given a description of the classification procedure once all the necessary elements – the classifying previsions and utility functions – are in place. We do not focus on how to choose appropriate utility functions, but use simple, reasonable ones: indicator functions, i.e., we let $U_{\mathfrak{c}} := I^{\mathfrak{c}}$. We do focus on the construction of the classifying previsions, which is done in two main steps:

(i) We first build a joint class-attribute model, a coherent lower prevision on $\mathscr{L}_{\mathfrak{C}\times\mathfrak{A}}$, that contains all information we know about the idealized process that generates classified objects, i.e., class-attribute pairs. A number of basic assumptions are made that form the skeleton of the class-attribute model and then its soft tissue is learned from a learning set of pre-classified objects, i.e., a finite sequence $\check{x}: (\mathfrak{C}\times\mathfrak{A})^*$ of class-attribute pairs.

(ii) The classifying previsions are then obtained as updated previsions of the class-attribute model, roughly speaking, by conditioning on the observed attribute tuple $\mathfrak{a}$ using natural or regular extension.

The details of both steps are worked out for a special subset of credal classifiers in the next subsection.

After learning a classifier, its performance should be validated, i.e., it must be checked if it is a good (enough) classifier. We are not going to discuss validation procedures in this thesis.

4.3.2 *The naive credal classifier*

Zaffalon's naive credal classifier [1999, 2001, 2002] is a credal classifier for categorical data: both the set of classes and the attribute sets must by good enough approximation be discrete and unordered. One can apply it to problems with attributes that take values in a continuous set by performing a discretization of the data. It is efficient and has been successfully applied to a number of problems [see, e.g., Zaffalon 2005; Zaffalon et al. 2003].

What makes a credal classifier naive? Any credal classifier is based on a class-attribute model for the possibility space $\mathfrak{C}\times\mathfrak{A}$. When the number of attributes is not small, this space is too big to be practical: it grows exponentially in the number of attributes; for example, adding a binary attribute doubles its size. And why is this not practical? Because the learning set (normally) does not grow exponentially in the number of attributes, but usually stays constant. So this means the relative number of different object types (attribute tuples) something is learned about sharply decreases with the number of attributes and therefore also the generalizing, thus classifying power of the model decreases. The obvious

solution is to add assumptions that increase the generalizing power of the class-attribute model. The assumption that is made in the naive credal classifier is that conditional on the class, the attributes are independent; this is a very strong and thus naive assumption indeed, as it is seldomly satisfied in practice. However, despite of this, the resulting classifier is surprisingly successful in practice. (Remark that if some specifics about dependencies between attributes are actually known, it would be even better to use this information to increase the generalizing power, if possible.)

Friedman [1997] and Domingos & Pazzani [1997] discuss the grounds for the surprising success of the naive Bayesian classifier.

Mathematically, naiveness implies that for every class $\mathfrak{c}$, there is a global attribute model, a coherent lower prevision $\underline{P}(\bullet|\mathfrak{c})$ on $\mathscr{L}_{\mathfrak{A}}$, that can be written as a product of individual attribute models (cf. §1.3.6$_{62}$). We are only going to consider the case of individual attribute models that can be easily specified as lower envelopes and – out of mathematical convenience – use a type-1 product (1.88)$_{64}$ to combine them. Let J be the finite index set describing the different components of the attribute tuple, so $\mathfrak{A} := \bigtimes_{j:J}\mathfrak{A}_j$. For every class $\mathfrak{c}$ and for each j in J, we have an individual attribute model, a coherent lower prevision $\underline{P}_j(\bullet|\mathfrak{c})$ on $\mathscr{L}_{\mathfrak{A}_j}$ that can be written as a lower envelope of a set $\{P_{\gamma_j}(\bullet|\mathfrak{c}) \mid \gamma_j : \Gamma_j\}$ of linear previsions on $\mathscr{L}_{\mathfrak{A}_j}$.

Let $\Gamma := \bigtimes_{j:J}\Gamma_j$; by the naiveness assumption, the per-class global model can then be written as an independent lower envelope (1.88)$_{64}$:

$$\underline{P}(\bullet|\mathfrak{c}) = \inf_{\gamma:\Gamma}\bigtimes_{j:J}P_{\gamma_j}(\bullet|\mathfrak{c}). \tag{4.47}$$

We collect these per-class global attribute models in an attribute model, a conditional lower prevision (cf. (1.59)$_{54}$ and (1.60)$_{54}$)

$$\underline{P}(\bullet|\mathfrak{C}) := \sum_{\mathfrak{c}:\mathfrak{C}}\underline{P}(\bullet|\mathfrak{c})\cdot I^{\mathfrak{c}}. \tag{4.48}$$

We separately learn each of the per-class individual attribute models from the learning set. We also learn a class model, a coherent lower prevision $\underline{P}$ on $\mathscr{L}_{\mathfrak{C}}$. What this learning entails practically will become clear when we turn attention to concrete attribute and class models (in a paragraph or three).

Because the class model $\underline{P}$ is then (trivially) only specified on gambles that are measurable with respect to the partition of the conditional lower prevision $\underline{P}(\bullet|\mathfrak{C})$ that is the attribute model, we can use marginal extension (cf. §1.3.5$_{61}$) to calculate their natural extension, the class-attribute model

$$\underline{E} := \underline{P}\big(\underline{P}(\bullet|\mathfrak{C})\big). \tag{4.49}$$

With the structure of the naive credal classifier's class-attribute model in place, we can make things more concrete: First of all, the class model $\underline{P}$ and the per-class individual attribute models $\underline{P}(\bullet|\mathfrak{c})$ all need to say something about concrete things, the unknown class and attributes of one

unobserved object, and not about any parameter that describes the process that generates these objects. So they are all immediate predictive lower previsions.

Given the fact that both the classes as well as each of the individual attributes are restricted to categorical variables, using IDMM's $(3.101)_{135}$ is an option we are already familiar with. It is also the option taken by Zaffalon [2002].

Plugging $(3.101)_{135}$ into (4.47)–$(4.49)_{\frown}$ results in

$$\begin{aligned} \underline{E} &:= \underline{\mathrm{Wa}}\big(\textstyle\sum_{\mathfrak{c}:\mathfrak{C}} I^{\mathfrak{c}} \cdot \underline{P}(\bullet|\mathfrak{c}) \,\big|\, \check{m}, s\big), \quad \text{with for each } \mathfrak{c} \text{ in } \mathfrak{C} \\ \underline{P}(\bullet|\mathfrak{c}) &:= \min\big\{\textstyle\bigtimes_{j:J} \mathrm{Wa}(\bullet|\check{m}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}} \cdot t_j) \,\big|\, t : \bigtimes_{j:J} \Delta_{\mathfrak{A}_j}\big\}, \end{aligned} \tag{4.50}$$

where

(i) s is the strictly positive real number of pseudocounts chosen for the class model,

(ii) $\check{m}$ is the count vector for the classes observed in the learning set (which can be thought of as a sequence $\check{x}$ in $(\mathfrak{C} \times \mathfrak{A})^*$), with $\check{n} := \sum \check{m}$,

(iii) $s_{j|\mathfrak{c}}$ is the strictly positive real number of pseudocounts chosen for the j-th individual attribute model contingent on class $\mathfrak{c}$,

(iv) $\check{m}_{j|\mathfrak{c}}$ is the count vector for j-th attribute of those objects in the learning set that belong to the class $\mathfrak{c}$, with $\check{m}_{\mathfrak{c}} = \check{n}_{j|\mathfrak{c}} := \sum \check{m}_{j|\mathfrak{c}}$.

Notice that we do not need to pay much attention to the learning aspect here, as we can just pick the appropriate inference model from the toolbox we have built up.

Let us include a toy example of how a learning set relates to the different count vectors mentioned above, to clarify and complement the explanation just given: Consider a classification problem with a class set $\mathfrak{C} := \{\text{dead}, \text{alive}\}$ and two attributes, citric acid response, with attribute set $\mathfrak{a}_{\mathrm{a}} := \{\text{contraction}, \text{still}\} = \{\mathrm{c}, \mathrm{s}\}$ and a time on ice (in hours), with attribute set $\mathfrak{A}_{\mathrm{h}} := \{<10, [10,20[, \geq 20\}$; it concerns oysters. The learning set is given in the following table:

vitality	citric acid response	time on ice
dead	still	<10
dead	still	≥ 20
alive	contraction	<10
dead	contraction	$[10,20[$
alive	contraction	$[10,20[$
dead	still	≥ 20

It corresponds to the following count vectors:

$$\begin{aligned} \check{m} &:= (\check{m}_{\text{dead}}, \check{m}_{\text{alive}}) = (4,2), \\ \check{m}_{\text{a|dead}} &:= \big((\check{m}_{\text{a|dead}})_{\mathrm{c}}, (\check{m}_{\text{a|dead}})_{\mathrm{s}}\big) = (1,3), \end{aligned}$$

$$\check{m}_{\text{a|alive}} \coloneqq \big((\check{m}_{\text{a|alive}})_{\text{c}}, (\check{m}_{\text{a|alive}})_{\text{s}}\big) = (2,0),$$
$$\check{m}_{\text{h|dead}} \coloneqq \big((\check{m}_{\text{h|dead}})_{<10}, (\check{m}_{\text{h|dead}})_{[10,20[}, (\check{m}_{\text{h|dead}})_{\geq 20}\big) = (1,1,2),$$
$$\check{m}_{\text{h|alive}} \coloneqq \big((\check{m}_{\text{h|alive}})_{<10}, (\check{m}_{\text{h|alive}})_{[10,20[}, (\check{m}_{\text{h|dead}})_{\geq 20}\big) = (1,1,0).$$

In (4.50), we still have the freedom to choose the values for the different pseudocounts. If these parameters are interpreted as just determining the learning speed, a logical choice would be to keep them constant and independent of each other. However, if they are interpreted as a number of hypothetical samples, it would be logical to impose, for each j, the constraint $s = \sum_{\mathfrak{c}:\mathfrak{C}} s_{j|\mathfrak{c}}$; i.e., the hypothetical samples are distributed among the classes. (Recall that when we were learning Markov chains in §3.4.4$_{149}$, we also encountered this issue and discussed it between (3.127)$_{150}$ and (3.128)$_{151}$.) The former option leads to easier optimization problems, but here we choose the latter path to stay in line with Zaffalon [2001]; however, the former path will be chosen in the next subsection, where keeping the optimization problems simple is far more critical. Zaffalon [2001] distributes the hypothetical sample according to the class model, so (4.50) becomes

$$\begin{aligned} \underline{E} &= \min\nolimits_{r:\Delta_{\mathfrak{C}}} \mathrm{Wa}\big(\textstyle\sum_{\mathfrak{c}:\mathfrak{C}} I^{\mathfrak{c}} \cdot \underline{P}(\bullet|\mathfrak{c}) \,\big|\, \check{m} + s\cdot r\big), \quad \text{with for each } \mathfrak{c} \text{ in } \mathfrak{C} \\ \underline{P}(\bullet|\mathfrak{c}) &= \min\big\{\textstyle\bigtimes_{j:J} \mathrm{Wa}(\bullet|\check{m}_{j|\mathfrak{c}} + s\cdot r_{\mathfrak{c}}\cdot t_j) \,\big|\, t:\bigtimes_{j:J}\Delta_{\mathfrak{A}_j}\big\}. \end{aligned} \tag{4.51}$$

Here, the per-class attribute models cannot be seen separately from the class model.

Now, to classify an object with attribute tuple $\mathfrak{a}$, we need to be able to decide, for each pair of classes $\mathfrak{d}$ and $\mathfrak{e}$, whether or not $\underline{E}(I^{\mathfrak{d}} - I^{\mathfrak{e}}|\mathfrak{a}) > 0$ (the maximality decision criterion).

(i) To calculate the updated prevision in case $\underline{E}(\mathfrak{C} \times \imath\mathfrak{a}) > 0$, both natural extension (1.83)$_{61}$ and regular extension (1.70)$_{58}$ give

$$\underline{E}(I^{\mathfrak{d}} - I^{\mathfrak{e}}|\mathfrak{a}) \propto \underline{E}\big((I^{\mathfrak{d}} - I^{\mathfrak{e}})\cdot I^{\mathfrak{a}}\big),$$

with the positive proportionality constant $1/\underline{E}(\mathfrak{C} \times \imath\mathfrak{a})$. So we then have $\underline{E}\big((I^{\mathfrak{d}} - I^{\mathfrak{e}})\cdot I^{\mathfrak{a}}\big) > 0$ as a criterion.

(ii) To calculate the regular extension in case $\underline{E}(\mathfrak{C} \times \imath\mathfrak{a}) = 0$, some linear previsions have to be omitted from $\mathscr{M}\underline{E}$, but doing so cannot change the sign of the criterion. We can therefore keep these previsions.

(iii) In case $\underline{E}(\mathfrak{C}\times\imath\mathfrak{a}) = 0$, the natural extension and, in case $\overline{E}(\mathfrak{C}\times\imath\mathfrak{a}) = 0$, also the regular extension is the vacuous lower prevision $\underline{P}^{\mathfrak{C}}$; so then $\underline{E}(I^{\mathfrak{d}} - I^{\mathfrak{e}}|\mathfrak{a}) = -1 < 0$ for any pair $(\mathfrak{d},\mathfrak{e})$, which means that no class is undominated and thus all classes are incomparable.

Zaffalon [2001] implicitly uses regular extension and as the IDMM by construction never assigns upper probability zero to any event, he can always use the criterion $\underline{E}\big((I^{\mathfrak{d}} - I^{\mathfrak{e}})\cdot I^{\mathfrak{a}}\big) > 0$. Using (4.51) and (3.90)$_{128}$

then leads us to checking the sign of

$$
\begin{aligned}
&\underline{E}\big((I^{\mathfrak{d}} - I^{\mathfrak{e}}) \cdot I^{\mathfrak{a}}\big) \\
&\quad = \min_{r:\Delta_{\mathfrak{C}}} \Big(\tfrac{\check{m}_{\mathfrak{d}} + s \cdot r_{\mathfrak{d}}}{\check{n} + s} \cdot \min\Big\{ \textstyle\prod_{j:J} \frac{(\check{m}_{j|\mathfrak{d}})_{\mathfrak{a}_j} + s \cdot r_{\mathfrak{d}} \cdot (t_j)_{\mathfrak{a}_j}}{\check{n}_{j|\mathfrak{d}} + s \cdot r_{\mathfrak{d}}} \;\Big|\; t : \textstyle\bigtimes_{j:J} \Delta_{\mathfrak{A}_j} \Big\} \\
&\qquad\qquad - \tfrac{\check{m}_{\mathfrak{e}} + s \cdot r_{\mathfrak{e}}}{\check{n} + s} \cdot \max\Big\{ \textstyle\prod_{j:J} \frac{(\check{m}_{j|\mathfrak{e}})_{\mathfrak{a}_j} + s \cdot r_{\mathfrak{e}} \cdot (t_j)_{\mathfrak{a}_j}}{\check{n}_{j|\mathfrak{e}} + s \cdot r_{\mathfrak{e}}} \;\Big|\; t : \textstyle\bigtimes_{j:J} \Delta_{\mathfrak{A}_j} \Big\} \Big) \\
&\quad \propto \min_{r:\Delta_{\mathfrak{C}}} \Big((\check{m}_{\mathfrak{d}} + s \cdot r_{\mathfrak{d}}) \cdot \textstyle\prod_{j:J} \frac{(\check{m}_{j|\mathfrak{d}})_{\mathfrak{a}_j}}{\check{n}_{j|\mathfrak{d}} + s \cdot r_{\mathfrak{d}}} - (\check{m}_{\mathfrak{e}} + s \cdot r_{\mathfrak{e}}) \cdot \textstyle\prod_{j:J} \frac{(\check{m}_{j|\mathfrak{e}})_{\mathfrak{a}_j} + s \cdot r_{\mathfrak{e}}}{\check{n}_{j|\mathfrak{e}} + s \cdot r_{\mathfrak{e}}} \Big) \\
&\quad = \min_{r:\Delta_{\mathfrak{C}}} \Big((\check{m}_{\mathfrak{d}} + s \cdot r_{\mathfrak{d}})^{1-|J|} \cdot \textstyle\prod_{j:J} (\check{m}_{j|\mathfrak{d}})_{\mathfrak{a}_j} \\
&\qquad\qquad - (\check{m}_{\mathfrak{e}} + s \cdot r_{\mathfrak{e}})^{1-|J|} \cdot \textstyle\prod_{j:J} \big((\check{m}_{j|\mathfrak{e}})_{\mathfrak{a}_j} + s \cdot r_{\mathfrak{e}} \big) \Big);
\end{aligned}
$$

if we fix $r_{\mathfrak{d}}$ and $r_{\mathfrak{e}}$ such that $r_{\mathfrak{d}} + r_{\mathfrak{e}} \in [0, 1[$, then we can lower the function's value by increasing one of them ($r_{\mathfrak{e}}$ for $|J| \leq 1$, $r_{\mathfrak{d}}$ for $|J| > 1$), so we may assume without loss of generality that $r_{\mathfrak{d}} + r_{\mathfrak{e}} = 1$:

$$
\begin{aligned}
&= \min_{r:[0,1]} \Big((\check{m}_{\mathfrak{d}} + s \cdot (1 - r))^{1-|J|} \cdot \textstyle\prod_{j:J} (\check{m}_{j|\mathfrak{d}})_{\mathfrak{a}_j} \\
&\qquad\qquad - (\check{m}_{\mathfrak{e}} + s \cdot r)^{1-|J|} \cdot \textstyle\prod_{j:J} \big((\check{m}_{j|\mathfrak{e}})_{\mathfrak{a}_j} + s \cdot r \big) \Big).
\end{aligned}
$$

This last, one-dimensional optimization problem can be solved straightforwardly [Zaffalon 2001].

We have now familiarized ourselves enough with the structure and important aspects of the naive credal classifier to be able to generalize it to individual attribute models (and perhaps even class models) that are not restricted to the IDMM, i.e., to categorical data. This is the topic of the next subsection.

4.3.3 *Generalizing the naive credal classifier*

Quaeghebeur & De Cooman [2005] and Quaeghebeur et al. [2005] give earlier reports on the ideas presented in this subsection.

In the previous subsection, we became really committed to learning from categorical data in the paragraph preceding $(4.50)_{184}$. Let us go back one step, to (4.47)–$(4.49)_{183}$, which presents the structure of the class-attribute model after imposing naiveness. Of course, instead of using IDMMs for the class model and individual attribute models, other immediate predictive lower previsions can be plugged in.

For example, so-called nonparametric immediate predictive inference models could be used as either a drop-in replacement for the IDMMs used [Coolen & Augustin 2009], or even to model one-dimensional continuous attributes [Augustin & Coolen 2004]. These models have no tuning parameter that characterizes the speed of learning. In this, they are very different from the IDMM, where the pseudocounts parameter takes this role. Apart from post-data exchangeability, these models also make no explicit assumptions about the process generating the objects and their attributes. In this, they are very different from the class of infer-

ence models we have introduced in §4.2.1$_{173}$, which are meant for situations where the object's attributes are assumed to be generated by some known exponential family sampling model with unknown parameter (and which are otherwise completely characterized by pseudocounts).

Although it is good to realize that other, less specific options are available, it is of course this last group of inference models – the immediate predictive variants (4.43)$_{174}$, to be precise – we are going to use here to generalize the naive credal classifier to noncategorical attributes. Replacement of the class model by a noncategorical one (to one for, e.g., Poisson sampling or normal sampling) would transform the classification problem we started out with into a regression problem, which would lead us too far, so we stick to the assumption that the class model is indeed categorical and we keep on using an IDMM for it.

The first thing that we do is write down the general form of the class-attribute models for which the attribute models are now allowed to be immediate predictive previsions of the form (4.43)$_{174}$: we replace (4.50)$_{184}$ by

$$\begin{aligned} \underline{E} &:= \underline{\mathrm{Wa}}\big(\textstyle\sum_{\mathfrak{c}:\mathfrak{C}} I^{\mathfrak{c}} \cdot \underline{P}(\bullet|\mathfrak{c}) \,\big|\, \check{m}, s\big), \quad \text{with for each } \mathfrak{c} \text{ in } \mathfrak{C} \\ \underline{P}(\bullet|\mathfrak{c}) &:= \inf_{t:\times_{j:J}\mathscr{U}_j} \bigtimes_{j:J} \mathrm{Pf}^{\mathsf{a}_{j|\mathfrak{c}},\tau_{j|\mathfrak{c}}}\Big(\bullet \,\Big|\, \check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}, \tfrac{\tau_{j|\mathfrak{c}}\check{x}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}\cdot t_j}{\check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}}\Big) \end{aligned} \tag{4.52}$$

where s, the $s_{j|\mathfrak{c}}$, and $\check{m}$ are defined as before, in (i)–(ii)$_{184}$, but now with the additional restriction that the pseudocounts parameters are all specified separately from one another. Furthermore,

(i) $\mathsf{a}_{j|\mathfrak{c}}$ and $\tau_{j|\mathfrak{c}}$ are the functions characterizing which exponential family sampling model is assumed to generate the j-th individual attribute contingent on class $\mathfrak{c}$ (cf. §4.1.1$_{155}$), and

(ii) $\check{x}_{j|\mathfrak{c}}$ is the sample sequence extracted from $\check{x}$ for the j-th attribute of those objects in the learning set that belong to the class $\mathfrak{c}$, with $\check{n}_{\mathfrak{c}} := \nu\check{x}_{j|\mathfrak{c}}$.

Let us return to the toy example of §4.3.2$_{182}$ and see what changes we have to make to the list of processed data, when instead of modeling the time on ice attribute as a categorical variable, we model it as being generated by a gamma sampling model (see §B.1.5$_{213}$ for more details on this exponential family) We do this for both the 'dead' and 'alive' case. The new learning set is given in the following table: (only the last column has been modified)

vitality	citric acid response	time on ice
dead	still	9
dead	still	27
alive	contraction	4
dead	contraction	17
alive	contraction	11
dead	still	32

Of course, we can keep the previously obtained class count vector $\check{m}$, and the count vectors $\check{m}_{\mathrm{a}|\mathrm{dead}}$, and $\check{m}_{\mathrm{a}|\mathrm{alive}}$ for the citric acid test (which are generated by $\tau_{\mathrm{a}} := C_{\{\mathrm{c},\mathrm{s}\}}$). For the time on ice attribute, however, we have to replace the count vectors $\check{m}_{\mathrm{h}|\mathrm{dead}}$ and $\check{m}_{\mathrm{h}|\mathrm{alive}}$ by the mean sufficient statistics for the gamma model, whose generating function for both the 'dead' and 'alive' case is given by $\tau_{\mathrm{h}} := z : \mathbb{R}_{>0} ; (z, \ln z)$ (cf $(\mathrm{B}.27)_{213}$):

$$\tau_{\mathrm{h}} \check{x}_{\mathrm{h}|\mathrm{dead}} = \tfrac{1}{\check{n}_{\mathrm{h}|\mathrm{dead}}} \cdot \big(\textstyle\sum \check{x}_{\mathrm{h}|\mathrm{dead}}, (\sum \circ \ln) \check{x}_{\mathrm{h}|\mathrm{dead}}\big) \approx (21.25, 2.95),$$
$$\tau_{\mathrm{h}} \check{x}_{\mathrm{h}|\mathrm{alive}} = \tfrac{1}{\check{n}_{\mathrm{h}|\mathrm{alive}}} \cdot \big(\textstyle\sum \check{x}_{\mathrm{h}|\mathrm{alive}}, (\sum \circ \ln) \check{x}_{\mathrm{h}|\mathrm{alive}}\big) \approx (7.50, 1.89).$$

Now, as before, to classify an object with attribute tuple $\mathfrak{a}$, we need to be able to decide, for each pair of classes $\mathfrak{d}$ and $\mathfrak{e}$, whether or not $\underline{E}(I^{\mathfrak{d}} - I^{\mathfrak{e}}|\mathfrak{a}) > 0$ (the maximality decision criterion). However, if any of the attributes has a continuous sampling model, then in general $\overline{E}(\mathfrak{C} \times \imath\mathfrak{a}) = 0$. In this case, updating on $\imath\mathfrak{a}$ using natural and regular extension would result in a vacuous model and all classes would be incomparable. This vacuity can be circumvented by making the additional assumption that the values of the continuous attributes are idealizations of events with at least positive upper probability [Walley 1991, §6.10_{328}]; i.e., we make the assumption that the measurement procedure has a finite precision.

Updating then consists of first applying Bayes's rule for density functions [Walley 1991, §$6.10.4_{331}$] to each of the predictive linear product previsions (let $r : \Delta_{\mathfrak{C}}$ and $t : \bigtimes_{j:J} \mathscr{U}_j$)

$$E_{r,t} = \mathrm{Wa}\big(\textstyle\sum_{\mathfrak{c}:\mathfrak{C}} I^{\mathfrak{c}} \cdot P_t(\bullet|\mathfrak{c}) \;\big|\; \frac{\check{m}+s\cdot r}{\check{n}+s}\big), \quad \text{with for each } \mathfrak{c} \text{ in } \mathfrak{C}$$
$$P_t(\bullet|\mathfrak{c}) = \textstyle\bigtimes_{j:J} \mathrm{Pf}^{\mathfrak{a}_{j|\mathfrak{c}}, \tau_{j|\mathfrak{c}}}\big(\bullet \;\big|\; \check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}, \tfrac{\tau_{j|\mathfrak{c}} \check{x}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}} \cdot t_j}{\check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}}\big)$$

that define the class-attribute model $\underline{E} = \inf_{r:\Delta_{\mathfrak{C}}; t:\times_{j:J}\mathscr{U}_j} E_{r,t}$ and then taking the lower envelope of the resulting updated linear previsions [Walley 1991, §$6.10.4–6_{331–333}$]:

$$\underline{E}(\bullet|\mathfrak{a}) := \inf_{\substack{r:\Delta_{\mathfrak{C}} \\ t:\times_{j:J}\mathscr{U}_j}} \tfrac{1}{P_r(p_t(\mathfrak{a}|\mathfrak{C}))} \cdot P_r\big(\bullet \cdot p_t(\mathfrak{a}|\mathfrak{C})\big),$$

where $\underline{P} = \inf_{r:\Delta_{\mathfrak{C}}} P_r$ is the class model and $p_t(\bullet|\mathfrak{C})$ is the conditional probability density or mass function corresponding to one of the linear previsions $P_t(\bullet|\mathfrak{C})$ defining attribute model $\underline{P}(\bullet|\mathfrak{C}) = \inf_{t:\times_{j:J}\mathscr{U}_j} P_t$ (cf. $(4.48)_{183}$).

Bayes's rule for density functions can be applied without worries, as the predictive probability density or mass function $(4.28)_{167}$ for exponential families is strictly positive on its domain for all allowed parameter values. So in particular, we know that

$$P_r(p_t(\mathfrak{a}|\mathfrak{C})) = \textstyle\sum_{\mathfrak{c}:\mathfrak{C}} \frac{\check{m}_{\mathfrak{c}} + s\cdot r_{\mathfrak{c}}}{\check{n}+s} \cdot \prod_{j:J} p_{t_j}(\mathfrak{a}_j|\mathfrak{c}) > 0,$$

where for each $\mathfrak{c}$ in $\mathfrak{C}$ we have used the shorthand

$$p_{t_j}(\mathfrak{a}_j|\mathfrak{c}) := \mathrm{pf}^{\mathfrak{a}_{j|\mathfrak{c}},\tau_{j|\mathfrak{c}}}\left(\mathfrak{a}_j \,\middle|\, \check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}, \frac{\tau_{j|\mathfrak{c}}\check{x}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}\cdot t_j}{\check{n}_{j|\mathfrak{c}} + s_{j|\mathfrak{c}}}\right). \tag{4.53}$$

(Cf. $(4.28)_{167}$ for the expression of the right-hand side probability density or mass.) Also, the sign of $\underline{E}(\bullet|\mathfrak{a})$ and $\inf_{r:\Delta_{\mathfrak{C}};t:\times_{j:J}\mathscr{U}_j} P_r\big(\bullet\cdot p_t(\mathfrak{a}|\mathfrak{C})\big)$ is the same, so we can replace the criterion by

$$\inf_{r:\Delta_{\mathfrak{C}};t:\times_{j:J}\mathscr{U}_j} P_r\big((I^{\mathfrak{d}} - I^{\mathfrak{e}})\cdot p_t(\mathfrak{a}|\mathfrak{C})\big) > 0,$$

whose left-hand side can be rewritten as follows:

$$\begin{aligned}
&\min_{r:\Delta_{\mathfrak{C}}}\Big(\tfrac{\check{m}_{\mathfrak{d}}+s\cdot r_{\mathfrak{d}}}{\check{n}+s}\cdot\inf_{t:\times_{j:J}\mathscr{U}_j}\textstyle\prod_{j:J} p_{t_j}(\mathfrak{a}_j|\mathfrak{d}) \\
&\qquad - \tfrac{\check{m}_{\mathfrak{e}}+s\cdot r_{\mathfrak{e}}}{\check{n}+s}\cdot\sup_{t:\times_{j:J}\mathscr{U}_j}\textstyle\prod_{j:J} p_{t_j}(\mathfrak{a}_j|\mathfrak{e})\Big) \\
\propto\ &\min_{r:\Delta_{\mathfrak{C}}}\Big((\check{m}_{\mathfrak{d}}+s\cdot r_{\mathfrak{d}})\cdot\textstyle\prod_{j:J}\inf_{t_j:\mathscr{U}_j} p_{t_j}(\mathfrak{a}_j|\mathfrak{d}) \\
&\qquad - (\check{m}_{\mathfrak{e}}+s\cdot r_{\mathfrak{e}})\cdot\textstyle\prod_{j:J}\sup_{t_j:\mathscr{U}_j} p_{t_j}(\mathfrak{a}_j|\mathfrak{e})\Big);
\end{aligned}$$

or, considering the expression between parentheses is linear in the components of r,

$$= \check{m}_{\mathfrak{d}}\cdot\textstyle\prod_{j:J}\inf_{t_j:\mathscr{U}_j} p_{t_j}(\mathfrak{a}_j|\mathfrak{d}) - (\check{m}_{\mathfrak{e}}+s)\cdot\prod_{j:J}\sup_{t_j:\mathscr{U}_j} p_{t_j}(\mathfrak{a}_j|\mathfrak{e}).$$

This last expression is the simplest one we can obtain without specifying the individual per-class attribute models.

Our decision criterion for ordering the classes consists of checking whether this last expression is strictly positive. The computational complexity of doing this is determined by the $2\cdot|J|$ uncoupled optimization problems that appear. For some exponential families – most notably for multinomial sampling – analytical solutions can be found for these optimization problems, for the others – such as for gamma sampling §B.1.5$_{213}$ – recourse to numerical methods is necessary. Although I expect that these optimization problems will not pose insurmountable challenges for most common exponential families, I cannot at this point, without having investigated the behavior of $(4.28)_{167}$ for all these families, make any more definite statements.

As regards the quality of the resulting classifier: for this we need to wait for an actual implementation, which unfortunately does not exist yet. This is also the reason no numerical example has been included here.

CONCLUSIONS

```
karma police
arrest this man
he talks in maths
he buzzesLikeAfridge
hes like a detuned radio.
```

Radiohead [1997]

Dear reader, if you have worked your way mostly linearly through all or a large part of the pages preceding this one: I salute you. To those who landed here after a nonlinear or inversely linear flight: welcome.

In these conclusions, I will reflect upon the topics presented in the four main chapters of this thesis. I will highlight my contributions and muse on the question "What next?"

5.0.4 *Modeling uncertainty*

In Chapter 1, 'Modeling uncertainty'$_{28}$, I presented an overview of a large part of the basic theory of coherent lower previsions. The primary intention of this chapter was to serve as a basis for the rest of the thesis. However, if it had only served that purpose, I could have achieved the same goal by extensively citing the literature (read: Walley's [1991] book).

One other, selfish reason was that writing things out oneself can be extremely enlightening: I understand the basic theory much better now. I also wished to present the theory slightly differently, with the hope of making it more accessible, or at least serve as inspiration for later accessible introductions, for which there is a genuine need. The main difference is the stronger focus on desirability as a basis for defining both unconditional and conditional previsions and working more explicitly with sets of desirable gambles [this focus is influenced by others, e.g., De Cooman & Miranda 2007, §2]; the emphasis on marginally desirable gambles as a link between the two models is the most personal touch. An important advantage of sets of desirable gambles is that they allow for simple, intuition-building graphical illustrations of many concepts in the theory of imprecise probabilities.

I believe that well thought out graphical illustrations are *very* useful explanatory tools.They should be used far more often and writers should be encouraged to learn the necessary tools and skills.

From my experience with writing this chapter, working with and thinking in terms of sets of desirable gambles can be very fruitful for obtaining theoretical and perhaps even practical results. I think that desirability can serve as an excellent basis for an accessible basic introduction to the techniques of the theory of imprecise probabilities.

Previsions and credal sets would then be introduced as derived concepts. The treatment should be restricted to finite possibility spaces, as infinities often only create technical problems and provide little or no additional insight.

5.0.5 *Extreme lower probabilities*

In Chapter 2, 'Extreme lower probabilities'$_{66}$, I gave an overview of the current state-of-the-art knowledge about extreme lower probabilities on finite possibility spaces. In most cases, the sets of lower probabilities that satisfy some property form a convex polytope in a – with regard to the size of the possibility space – high-dimensional space. The extreme points of these polytopes, which provide a complete characterization, are the extreme lower probabilities. Most interesting properties lower probabilities can possess are expressed using sets of constraints, which more or less correspond to the polytope's faces. To pass from constraints to extreme points, vertex enumeration algorithms are used.

Although vertex enumeration is a commonly used technique, my contribution lies in applying it to this particular problem, for which it was first necessary to generate a manageable set of constraints for the different properties of interest. The computer program I wrote allowed the discovery of many hitherto unknown sets of extreme lower probabilities for a lot of properties. I also gave an analysis of a number of interesting cases, wherever possible guided by the adjacency graph of the set of lower probabilities under scrutiny. Some regularities were uncovered by this analysis, which I continued in the Herbarium$_{194}$, but the most striking fact was the combinatorial explosion of the number of extreme coherent lower probabilities with the size of the possibility space.

This combinatorial explosion, which limits the practicality of calculating and thus working with extreme lower probabilities (it was our original hope to use extreme lower probabilities for approximation purposes), leads me to believe that not much more can be learned by using more powerful computers to obtain results for larger possibility spaces. Rather, to gain more insight, and perhaps even to find a clue that indicates analytical results are possible, further research should focus on manageable possibility spaces, but move from lower probabilities to lower previsions; to wit, apart from indicators, add more and more other gambles into the mix and see how the set of extreme points changes with each gamble added.

5.0.6 *Inference models*

In Chapter 3, 'Inference models'$_{94}$, I reported on research – done in close collaboration with Gert de Cooman and Enrique Miranda – about exchangeable lower previsions and about immediate prediction under representation insensitivity. However, the wish to combine these two

subjects as seamlessly as possible, in combination with my fascination for marginally desirable gambles, led me to propose a novel concept of finite exchangeability for sets of desirable gambles, that – according to my taste – better captures the idea behind exchangeability. This allowed me to evade a number of bothersome technicalities that had previously prevented us from finding an important result: that the IDMM (IDM) is the unique inference model for categorical data that is (infinitely) exchangeable, representation insensitive and simple in the sense that it produces linear-vacuous immediate predictions, under the added condition, however, that one follows the constant hyperparameter path (a *new* bothersome technical requirement). To show how easily applicable the ID(M)M is, I decided to not just point to the literature for examples of applications, but present two which I have worked on myself: game-theoretic learning and learning of Markov chains.

Discovering how to derive the ID(M)M from first principles is in itself nice, but in some sense even nicer is the idea that led to the removal of the obstacles on the path to this discovery: a definition for the exchangeability of sets of desirable gambles. This idea can lead to a fully worked out theory of exchangeable sets of desirable gambles. (Gert de Cooman is making progress on this even as I write!) This idea could also be translated to other invariance assumptions. It would also be a very interesting challenge to see where one lands by not following the constant hyperparameter path. Concerning future work on the presented applications: the inference model for Markov chains needs to be tested in practice.

5.0.7 *Inference models for exponential families*

The ID(M)M's first appearance far preceded our derivation from first principles: it was introduced by Walley [1996] as a generalization of a group of classical, Bayesian inference models for categorical data and has been one of the most widely used imprecise-probabilistic inference models ever since. In Chapter 4, 'Inference models for exponential families'[154], we added another generalization step. Categorical sampling is described by a multinomial likelihood function and the ID(M)M is based on the conjugate family of Dirichlet distributions, for which the parameters are split into a part interpretable as a number of counts and a part interpretable as an average one-sample sufficient statistic. Now, the multinomial likelihood is part of a class of formally similar regular exponential family likelihoods. For all of these, introducing an inference model based on the conjugate family is as easy as it was with the ID(M)M. To boot, the parameters of these families can be split up the same way as was done with the ID(M)M.

The more I think about it, the more it seems to me that this generalization was waiting to happen ever since the ID(M)M appeared. Although

there is some merit in having read the right books and papers, realizing that the pieces fit, and getting the idea out in the wild, I still have qualms about the fact that I have not yet managed to put the proposed inference models into practice. If this line of research is to be continued, this is one of the first things that need to be remedied. To this end, I have worked out how this can be done for credal classification. The lack of a simple analytical expression for immediate prediction for many exponential families could be a bothersome, but not insurmountable problem.

It would be an interesting and very challenging research program to see if, apart from the ID(M)M, other imprecise-probabilistic inference models for exponential families could be derived from first principles. For starters, this would entail identifying the assumptions encoded in the likelihood function (both about invariances and about the domain's structure) and struggling with a number of limiting arguments (from finite to infinite sample sequences and perhaps also from finite to infinite possibility spaces). One element already seems to be in place: the inference models I have proposed are simple in the sense that they satisfy the posterior contamination prediction property, i.e., the immediate predictive lower prevision of a single sample sufficient statistic is a mixture of a linear prevision with a (near-)vacuous lower prevision. (And why not try to do all this in terms of desirable gambles?)

On the more practical side, it could be very interesting to follow Walley & Bernard's [1999] lead and derive predictive lower and upper cumulative distribution functions. These allow us to make comparisons with frequentist methods that produce confidence intervals. Such comparisons are probably essential to be able to communicate with the many statisticians in the field using these methods and convince them of the usefulness of imprecise-probabilistic methods.

HERBARIUM
A basket of extreme lower probabilities

> The staple source of carbohydrates in Ethiopia is *injera*, a large, pancake-shaped substance made from *tef*, a nutty-tasting grain that is unique to Ethiopia and comes in three varieties: white, brown and red.
>
> Briggs [2005, p. 101]

Coherent lower probabilities are a source of reasonable uncertainty models; they can be written as a convex combination of extreme lower probabilities (cf. 'Extreme lower probabilities'$_{66}$). These lower probabilities and their extreme building blocks come in many varieties, depending on the properties they have to satisfy.

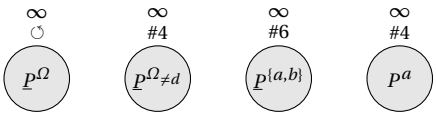
In this appendix, we include some additional sets of extreme lower probabilities as well as two technical lemmas necessary to derive the feasible definitions for avoiding sure loss (2.21)$_{79}$ and coherence (2.24)$_{82}$.

A.1 Selected extreme lower probabilities

In §2.3$_{85}$ of the chapter 'Extreme lower probabilities'$_{66}$, we presented and commented on several adjacency graphs and a list of extreme lower probabilities. In this section, we add one more list (containing the previously unknown extreme coherent lower probabilities on a possibility space of cardinality 4) and several adjacency graphs (for different cases involving permutation invariance). As in §2.3$_{85}$, all results are based on numerical output from my 'constraints' computer program.

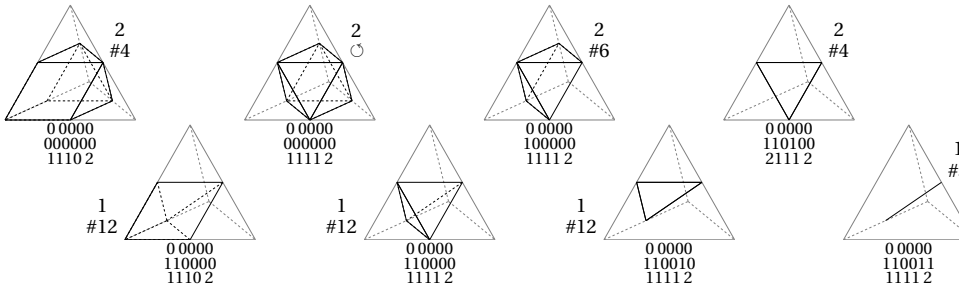
A.1.1 *Coherent on four*

We have discovered the 402 extreme coherent lower probabilities on a possibility space of 4 elements (let $\Omega := \{a, b, c, d\}$). They are listed below, grouped by the maximal denominator occurring in their components. As the adjacency graph for this case would be far too complex, this way of presenting things is an invitation to focus more on the combinatorial aspects and less on the geometrical ones.

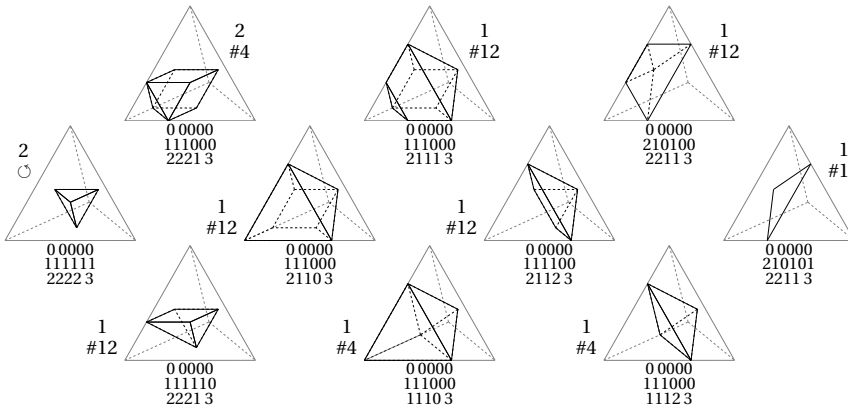


The first group (4 permutation classes) is the one formed by the 15 extreme coherent lower probabilities with maximal denominator 1. These are the vacuous lower probabilities, all of which are completely monotone. Only one is permutation invariant.

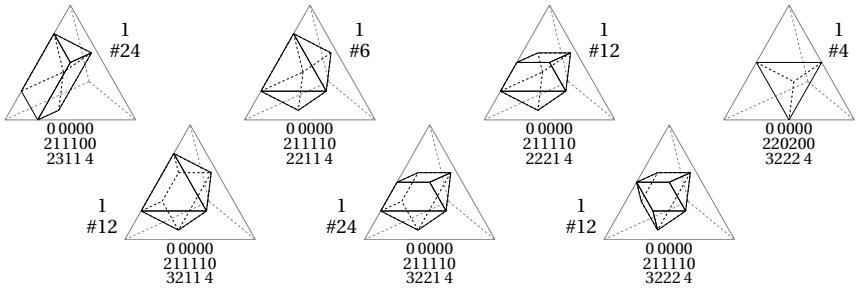
The second group (8 permutation classes) is formed by the 54 extreme coherent lower probabilities with maximal denominator 2. One is permutation invariant. Half of the classes are 2-monotone. Two classes have 2-dimensional credal sets and one has 1-dimensional credal sets.



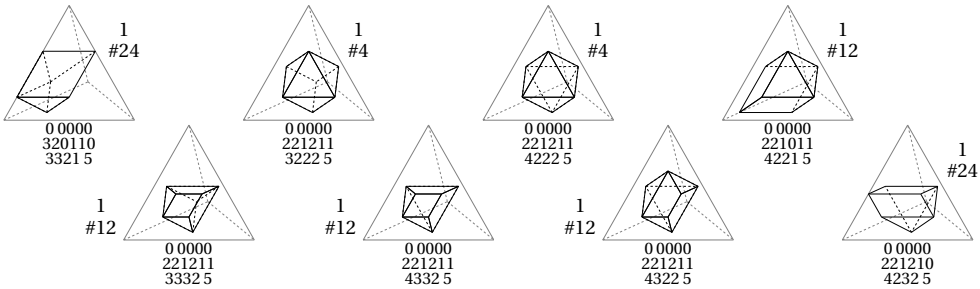
The third group (10 permutation classes) is formed by the 85 extreme coherent lower probabilities with maximal denominator 3. One is permutation invariant. Two of the classes are 2-monotone. One class has 2-dimensional credal sets.

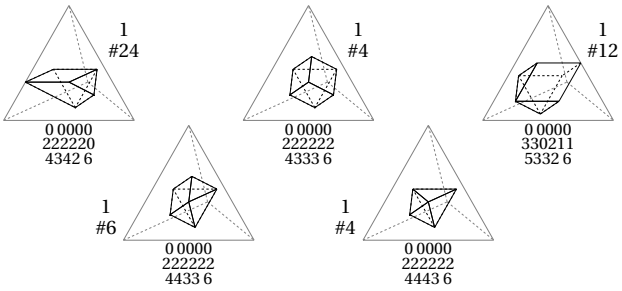


The fourth group (7 permutation classes) is formed by the 94 extreme coherent lower probabilities with maximal denominator 4:



The next-to-last group (8 permutation classes) is formed by the 104 extreme coherent lower probabilities with maximal denominator 5:



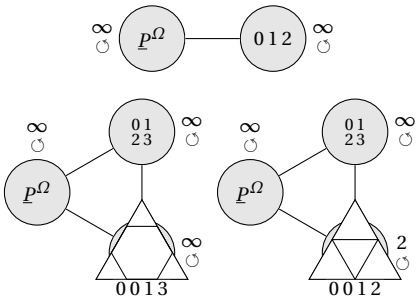


The last group (5 permutation classes) is formed by the 50 extreme coherent lower probabilities with maximal denominator 6.

None of the classes among the last three of the groups is 2-monotone, permutation invariant, or has non-3-dimensional credal sets.

A.1.2 *Staring at sub-cubical dice in various ways*

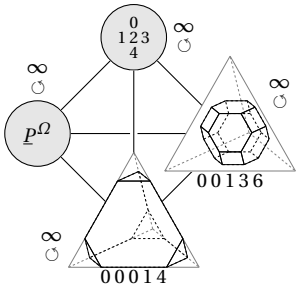
In §2.3.4$_{93}$, we looked at the adjacency graph of the set of permutation invariant 2-monotone lower probabilities on a possibility space of cardinality 6; we presented them as possible models for dealing with the uncertainty involved in betting with common six-faced cubical dice about which we have no face-specific information. In this subsection, we do the same, but for possibility spaces of lower cardinality; we combine permutation invariance with coherence or k-monotonicity (for each relevant k in $\mathbb{N}_{\geq 2}$).

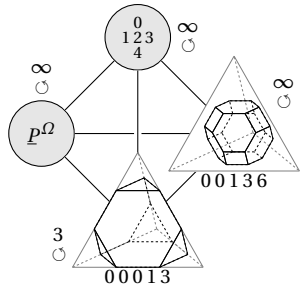


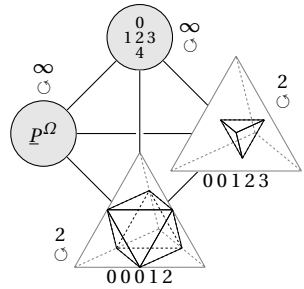
For $|\Omega| = 2$ (flipping or spinning a coin), we give the graph for the coherent case, which coincides with the completely monotone case.

For $|\Omega| = 3$ (e.g., tossing a very uncommon elongated triangular dipyramidic die, the downward facing side counts), we first give the completely monotone case and then the coherent case, which coincides with the 2-monotone case.

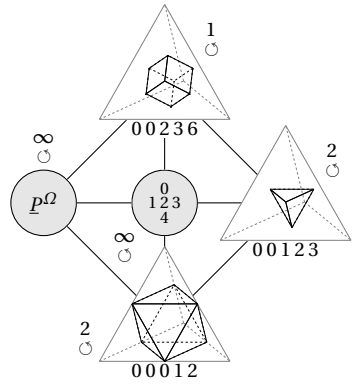
Unfolding an elongated triangular dipyramidic die:







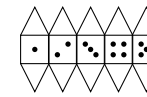
We give, for $|\Omega| = 4$ (tossing a less uncommon tetraedric die), respectively above from left to right, the completely monotone, 3-monotone, and 2-monotone cases and then, on the side, the coherent case. Notice that the first three adjacency graphs correspond to tetrahedra whose points can be written as a unique convex combination of the vertices, whereas for the coherent case, we find a pyramid with a quadrangular base and the additive extreme probability $0\ \frac{1}{4}\ \frac{1}{2}\ \frac{3}{4}\ 1$ in the apex. The decomposition is not unique anymore, for example:

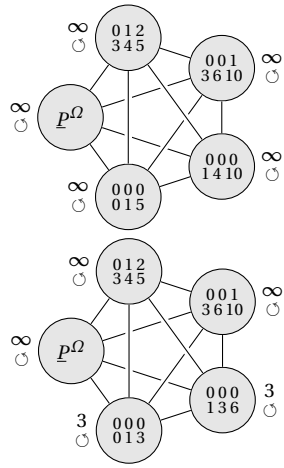


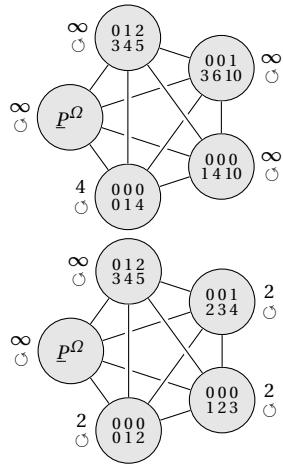
$$0\,0\,\tfrac{1}{4}\,\tfrac{1}{2}\,1 = \tfrac{3}{4}\cdot(0\,0\,\tfrac{1}{3}\,\tfrac{1}{2}\,1) + \tfrac{1}{4}\cdot(0\,0\,0\,\tfrac{1}{2}\,1)$$
$$= \tfrac{3}{4}\cdot(0\,0\,\tfrac{1}{3}\,\tfrac{2}{3}\,1) + \tfrac{1}{4}\cdot\underline{P}^{\Omega}.$$

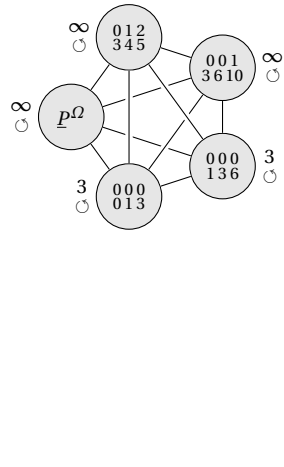
The situation for $|\Omega| = 5$ (e.g., an elongated pentagonal dipyramidic die) is similar. On the left we have the completely, 4-, 3-, and 2-monotone cases (pentachora, the 4-dimensional version of the tetrahedra); on the right the coherent case (also 4-dimensional, which can be seen by removing the central additive extreme lower probability $0\ \frac{1}{5}\ \frac{2}{5}\ \frac{3}{5}\ \frac{4}{5}\ 1$ and realizing the resulting graph corresponds to a 3-dimensional polytope).

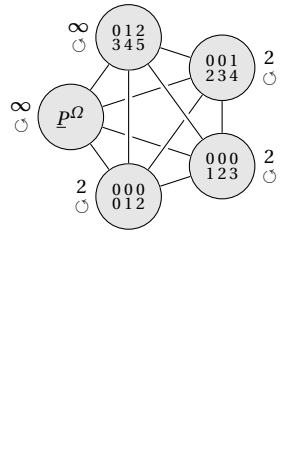
Unfolding an elongated pentagonal dipyramidic die:

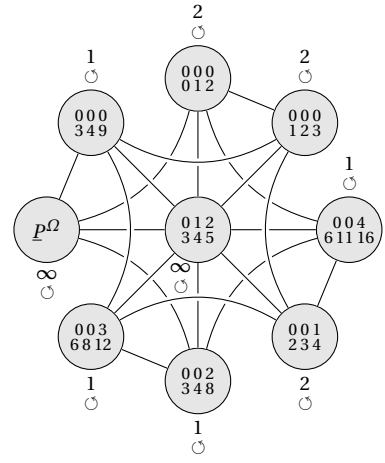












A.2 TECHNICAL LEMMAS

This section contains two technical lemmas used in the construction of feasible definitions for avoiding sure loss in §2.2.4$_{76}$ and for coherence in §2.2.5$_{80}$. For both, we assume that some finite possibility space Ω is given.

A.2.1 *Elimination of linear dependence*

This lemma is inspired by Walley [1991, §A1].

In this lemma, we consider a set of constraints on a fixed vector q in $\wp\Omega \to \mathbb{R}$. Each constraint is characterized by a set of events $\mathscr{C} :\subseteq \wp\Omega$ and coefficients λ in $(\mathbb{R}_{\neq 0})^{\mathscr{C}}$ such that $\sum_{C:\mathscr{C}} \lambda_C \cdot I^C = \rho$, where ρ is some fixed real number. The lemma states that by eliminating those sets $\mathscr{C}$ for which $\{I^C - qC \mid C:\mathscr{C}\}$ is linearly dependent, this set can be reduced to one that defines the same polyhedron. (Recall from §0.3.3$_{25}$ that for function application, subscripting is used for coefficients and postplacement is used otherwise.)

To state this lemma more formally, suppose $\wp\Omega$ is partitioned into $\mathscr{A}$ and $\mathscr{B}$, where $\mathscr{A}$ and $\mathscr{B}$ will correspond to the positive and negative components of λ, respectively. The lemma then gives two equivalent formulations for set of constraints on q we are interested in, the original one and the useful one, respectively:

$$\begin{aligned}&\forall \mathscr{C} :\subseteq \wp\Omega;\\ &\forall \lambda : (\mathbb{R}_{<0})^{\mathscr{C}\cap\mathscr{A}} \times (\mathbb{R}_{>0})^{\mathscr{C}\cap\mathscr{B}} \mathbin{\wedge} \textstyle\sum_{C:\mathscr{C}} \lambda_C \cdot I^C = \rho;\\ &\qquad \sup\{\textstyle\sum_{C:\mathscr{C}} \lambda_C \cdot (I^C - qC)\} \geq 0\end{aligned} \tag{A.1}$$

is equivalent to

$$\begin{aligned}&\forall \hat{\mathscr{C}} :\subseteq \wp\Omega \mathbin{\wedge} \neg \operatorname{dep}\{I^C - qC \mid C:\hat{\mathscr{C}}\};\\ &\forall \hat{\lambda} : (\mathbb{R}_{<0})^{\hat{\mathscr{C}}\cap\mathscr{A}} \times (\mathbb{R}_{>0})^{\hat{\mathscr{C}}\cap\mathscr{B}} \mathbin{\wedge} \big(\exists \hat{\rho} : \mathbb{R} ; \textstyle\sum_{C:\hat{\mathscr{C}}} \hat{\lambda}_C \cdot I^C = \hat{\rho}\big);\\ &\qquad \sup\{\textstyle\sum_{C:\hat{\mathscr{C}}} \hat{\lambda}_C \cdot (I^C - qC)\} \geq 0.\end{aligned} \tag{A.2}$$

(Readers less interested in technical derivations can immediately skip the rest of this subsection.)

To prove this equivalence, start from (A.1) and consider those $\mathscr{C}$ for which $\{I^C - qC \mid C:\mathscr{C}\}$ is linearly dependent. Then, following (2.19)$_{78}$, there is a μ in $\mathbb{R}^{\mathscr{C}}$ such that $\mu \neq (\mathscr{C} ; 0)$ and $\sum_{C:\mathscr{C}} \mu_C \cdot (I^C - qC) = 0$. So the sum appearing in the expression for these constraints can be rewritten as

$$\textstyle\sum_{C:\mathscr{C}} \lambda_C \cdot (1 + \gamma \cdot \frac{\mu_C}{\lambda_C}) \cdot (I^C - qC),$$

where the real number γ can be freely chosen.

Choosing $\gamma := -\lambda_D/\mu_D$, where D in $\mathscr{C}$ is such that $\mu_D \neq 0$, reduces the constraint to an equivalent one not explicitly dependent on D, which

can therefore be removed from $\mathscr{C}$; other C in $\mathscr{C}$, for which $1+\gamma\cdot\frac{\mu_C}{\lambda_C}=0$, can also be removed. Possibly after repeating this procedure, one obtains a replacement $\hat{\mathscr{C}}:\subset\mathscr{C}$ for which the set $\{I^C-qC\,|\,C:\hat{\mathscr{C}}\}$ is linearly independent.

For any C in $\mathscr{C}$, the signs of λ_C and $\hat{\lambda}_C:=\lambda_C\cdot(1+\gamma\cdot\frac{\mu_C}{\lambda_C})$ are the same when $0\le 1+\gamma\cdot\frac{\mu_C}{\lambda_C}=1-\frac{\lambda_D}{\mu_D}\cdot\frac{\mu_C}{\lambda_C}$. This can be guaranteed for all C in $\hat{\mathscr{C}}$ by choosing D such that $|\lambda_D/\mu_D|=\min_{C:\mathscr{C}\wedge\mu_C\neq 0}|\lambda_C/\mu_C|$, so the sign of the components of $\hat{\lambda}$ corresponds to the same partition $\{\mathscr{A},\mathscr{B}\}$ of $\wp\Omega$.

Remark that

$$\begin{aligned}\textstyle\sum_{C:\hat{\mathscr{C}}}\hat{\lambda}_C\cdot I^C&=\textstyle\sum_{C:\mathscr{C}}\lambda_C\cdot(1+\gamma\cdot\frac{\mu_C}{\lambda_C})\cdot I^C\\&=\textstyle\sum_{C:\mathscr{C}}(\lambda_C\cdot I^C+\gamma\cdot\mu_C\cdot qC)\\&=\hat{\rho}:=\rho+\textstyle\sum_{C:\mathscr{C}}\gamma\cdot\mu_C\cdot qC,\end{aligned}$$

so $\hat{\lambda}$ also satisfies a condition like the one given for λ, $\sum_{C:\mathscr{C}}\lambda_C\cdot I^C=\rho$. Also notice that ρ and $\hat{\rho}$ need not have the same sign.

A.2.2 *Preservation of linear independence*

In this lemma, we consider sets of functions of the type $\{I^C-qC\,|\,C:\mathscr{C}\}$, determined by some set of events $\mathscr{C}:\subseteq\wp\Omega$ and some vector q in $\mathscr{C}\to\mathbb{R}$. We show that if there are coefficients λ in $\mathbb{R}^{\mathscr{C}}$ such that $\sum_{C:\mathscr{C}}\lambda_C\cdot I^C=1$ (the 'assumption of the lemma', which also implies $\lambda\neq(\mathscr{C}\,;0)$), then linear (in)dependence of this set of functions is equivalent to linear (in)dependence of $\{I^C\,|\,C:\mathscr{C}\}$, in which q does not appear.

So, formally, what we want to prove is

$$\begin{aligned}&\forall\mathscr{C}:\subseteq\wp\Omega\wedge\big(\exists\lambda:\mathbb{R}^{\mathscr{C}}\,;\textstyle\sum_{C:\mathscr{C}}\lambda_C\cdot I^C=1\big);\\&\forall q:\mathscr{C}\to\mathbb{R};\\&\qquad\mathrm{dep}\{I^C-qC\,|\,C:\mathscr{C}\}\Leftrightarrow\mathrm{dep}\{I^C\,|\,C:\mathscr{C}\}.\end{aligned}\tag{A.3}$$

(Readers less interested in technical derivations can skip the rest of this subsection.)

First, we prove that linear dependence of $\{I^C-qC\,|\,C:\mathscr{C}\}$ implies linear dependence of $\{I^C\,|\,C:\mathscr{C}\}$: If $\{I^C-qC\,|\,C:\mathscr{C}\}$ is linearly dependent, then, following definition $(2.19)_{78}$, there is a μ in $\mathbb{R}^{\mathscr{C}}$ such that $\mu\neq 0$ and such that $\sum_{C:\mathscr{C}}\mu_C\cdot I^C=\sum_{C:\mathscr{C}}\mu_C\cdot qC$. Now define the map $c:=\mu:\mathbb{R}^{\mathscr{C}}\,;\sum_{C:\mathscr{C}}\mu_C\cdot qC$. Due to the assumption of the lemma,

$$\textstyle\sum_{C:\mathscr{C}}c\mu\cdot\lambda_C\cdot I^C=\sum_{C:\mathscr{C}}\mu_C\cdot qC=\sum_{C:\mathscr{C}}\mu_C\cdot I^C.$$

So with $\hat{\mu}:=\mu-c\mu\cdot\lambda$, we find that $\sum_{C:\mathscr{C}}\hat{\mu}_C\cdot I^C=0$ and thus we have shown that $\{I^C\,|\,C:\mathscr{C}\}$ is linearly dependent.

Now we prove that linear independence of $\{I^C-qC|C:\mathscr{C}\}$ implies linear independence of $\{I^C|C:\mathscr{C}\}$: If $\{I^C-qC|C:\mathscr{C}\}$ is linearly independent,

then for all μ in $\mathbb{R}^{\mathscr{C}}$ such that $\mu \neq 0$, it holds that

$$\sum_{C:\mathscr{C}} \mu_C \cdot I^C \neq \sum_{C:\mathscr{C}} \mu_C \cdot qC = c\mu.$$

Again due to the assumption of the lemma, we find it holds for all μ that $\sum_{C:\mathscr{C}} (\mu_C - c\mu \cdot \lambda_C) \cdot I^C \neq 0$. So if for all $\hat{\mu}$ in $\mathbb{R}^{\mathscr{C}}$ such that $\hat{\mu} \neq (\mathscr{C}\,;0)$ there is a μ such that $\hat{\mu} = \mu - c\mu \cdot \lambda$, then $\{I^C \,|\, C:\mathscr{C}\}$ is linearly independent and the lemma is proven.

To finish the proof, we have to show, under the given conditions on μ and $\hat{\mu}$, that the linear system $\hat{\mu} = \mu - c\mu \cdot \lambda$ of equations in μ can always be solved for μ when $\{I^C - qC \,|\, C:\mathscr{C}\}$ is linearly independent. Or, using the expression for $c\mu$ and considering that $\mu = 0$ implies $\hat{\mu} = 0$, it suffices to show that the system's coefficient matrix $\mathbb{1} - \lambda q^\mathsf{T}$ has a nonzero determinant. To do this, we create a well-chosen matrix, and rewrite it as two different products

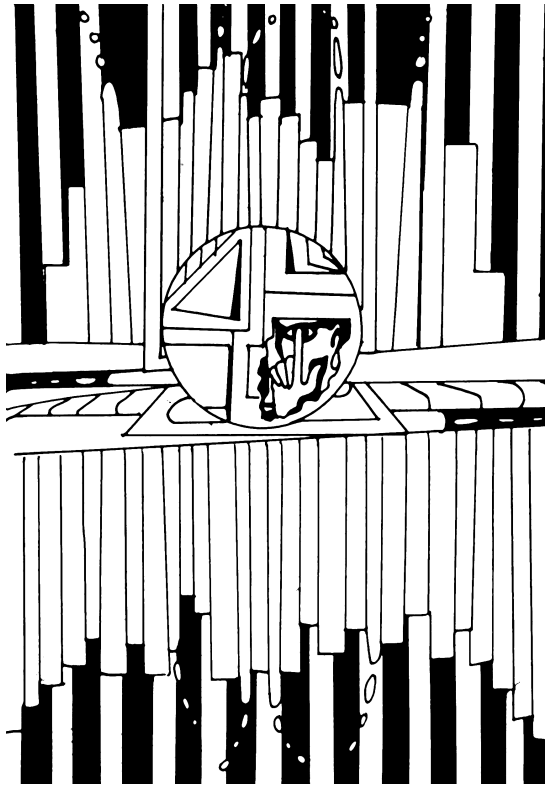
Matrix notation: identity matrix $\mathbb{1}$, zero vector $\mathbb{0}$, transposition $\bullet^\mathsf{T}$, dyadic product $\bullet\bullet^\mathsf{T}$, scalar product $\bullet^\mathsf{T}\bullet$, and determinant $|\bullet|$.

$$\begin{pmatrix} \mathbb{1} & \lambda \\ q^\mathsf{T} & 1 \end{pmatrix} = \begin{pmatrix} \mathbb{1} & \lambda \\ \mathbb{0}^\mathsf{T} & 1 \end{pmatrix} \begin{pmatrix} \mathbb{1} - \lambda q^\mathsf{T} & \mathbb{0} \\ q^\mathsf{T} & 1 \end{pmatrix} = \begin{pmatrix} \mathbb{1} & \mathbb{0} \\ q^\mathsf{T} & 1 \end{pmatrix} \begin{pmatrix} \mathbb{1} & \lambda \\ \mathbb{0}^\mathsf{T} & 1 - q^\mathsf{T}\lambda \end{pmatrix}.$$

Taking the determinant gives

$$\begin{vmatrix} \mathbb{1} & \lambda \\ q^\mathsf{T} & 1 \end{vmatrix} = |\mathbb{1} - \lambda q^\mathsf{T}| = 1 - q^\mathsf{T}\lambda.$$

So $|\mathbb{1} - \lambda q^\mathsf{T}| \neq 0$ when $1 \neq \sum_{C:\mathscr{C}} \lambda_C \cdot qC$. Or, because of the assumption of the lemma, when $\sum_{C:\mathscr{C}} \lambda_C \cdot I^C \neq \sum_{C:\mathscr{C}} \lambda_C \cdot qC$. This is satisfied because $\{I^C - qC \,|\, C:\mathscr{C}\}$ is linearly independent.



BESTIARIUM
A list of exponential families & their friends

> A cow's strategic and commercial sensibilities are not highly developed.
>
> Graves [1955, §58.6]

The list of regular exponential families one could come up with is uncountable, that much is clear from their description in §4.1.1$_{155}$: we can already construct a linear canonical exponential family for almost any sample space one can come up with. It is another question entirely which ones out of this panoply of families is practically useful as a sampling model. In 'Inference models for exponential families'$_{154}$ we have encountered two of them already: the normal and multi-category Bernoulli sampling models (in §4.1.2$_{159}$ and §4.1.3$_{160}$, respectively).

In this appendix, we have a look at some other exponential families that have proven their worth as sampling models. As we have done for the normal and multi-category Bernoulli sampling models, we derive the corresponding conjugate family and – for some – the immediate predictive family (cf. §4.1.6$_{168}$ and §4.1.7$_{171}$). For each, we also give a possible choice for the set of parameters that determines an ICEFM or IPEFM; but except for one family (§B.1.4$_{210}$), we do not illustrate the updating procedure (as was done in §4.2.2$_{178}$ and §4.2.3$_{179}$).

The material in this appendix is not original in the sense that the conjugate and predictive distributions derived here can be found elsewhere. What is more original is our focus on the parameterization in terms of the number of observations and their mean single-sample sufficient statistic.

This appendix can be of interest as a partial reference, but is written with the idea that it should allow the reader to get acquainted with a number of different exponential families and their friends. The treated exponential families are split up into those defined on continuous and discrete sample spaces – §B.1 and §B.2$_{215}$ respectively. Each exponential family is accorded its own subsection, the structure of which is always the same: first the family itself is presented and its characteristics are derived; then the form of its conjugate parametric prevision and its normalization factor are derived; for some, the immediate predictive prevision is found in a third step; each subsection always ends with a suggestion on how to choose a bounded convex set of single-sample sufficient statistics to build ICEFMs or IPEFMs starting from the parametric and predictive linear previsions.

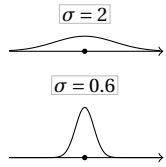
B.1 Continuous families

This section with exponential families on continuous sample spaces is divided in three parts. One of these parts, §B.1.1–§B.1.3$_{203–206}$, is inhabited by families that are a restriction or generalization of the normal family we have treated extensively in 'Inference models for exponential families'$_{154}$. Another, §B.1.4$_{210}$, contains the interesting von Mises family, which is mainly used to model planar directional sampling. The third, consisting of §B.1.5$_{213}$ and §B.1.6$_{214}$, is devoted to the gamma family and the exponential exponential family (sic), which are too common to be omitted.

B.1.1 *Centered normal sampling*

The centered normal family is the subset of the normal family we saw in §4.1.2$_{159}$ for which the mean is 0. It can be used whenever the mean μ in $\mathbb{R}$ is known; one just needs to do a coordinate translation over $-\mu$. Each member has $\mathscr{X} := \mathbb{R}$ as a sample space and is parameterized by its standard deviation, i.e., a $\phi := \sigma$ in $\Phi := \mathbb{R}_{>0}$. The centered normal linear prevision is $\mathrm{Nl}(\bullet|0,\sigma)$ (cf. $(4.9)_{159}$).

Two centered normal densities (plot restricted to $[-5,5]$): (the mean 0 is indicated with a dot)



Following the derivation in §4.1.2$_{159}$, we can easily derive the likelihood in exponential family form: let z be an observed real sample, then this likelihood can be written as

$$\begin{aligned}\tfrac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\Big(-\tfrac{1}{2}\cdot\big(\tfrac{z}{\sigma}\big)^2\Big) &= \tfrac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\big(-\tfrac{z^2}{2\cdot\sigma^2}\big)\\ &= \tfrac{1}{\sigma}\cdot\exp\big\langle -\tfrac{1}{2\cdot\sigma^2}\;\big|\; z^2\big\rangle\cdot\tfrac{1}{\sqrt{2\cdot\pi}}.\end{aligned}$$

Comparing this last expression with $(4.1)_{156}$, we see that the Euclidean vector space for this family has dimension $d := 1$ and that

$$\begin{aligned}&\tau := z:\mathbb{R}\,; z^2, \quad \text{so} \quad \mathscr{T} := \mathrm{co}\{z^2 \mid z:\mathbb{R}\} = \mathbb{R}_{\geq 0},\\ &\psi := \sigma:\mathbb{R}_{>0}\,; -\tfrac{1}{2\cdot\sigma^2}, \quad \text{so} \quad \Xi := \mathbb{R}_{<0},\end{aligned} \tag{B.1}$$

and also that

$$\mathsf{a} := \mathbb{R}\,; \tfrac{1}{\sqrt{2\cdot\pi}} \quad \text{and} \quad \mathsf{b} := \xi:\mathbb{R}_{<0}\,; \sqrt{-2\cdot\xi}. \tag{B.2}$$

The cumulant function for the centered normal family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi:\mathbb{R}_{<0}\,; -\tfrac{1}{2}\cdot\ln(-2\cdot\xi).$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find – letting $\xi := \psi\sigma$ – that

$$\mathrm{Nl}(\tau|0,\sigma) = (\nabla\kappa)\xi = -\tfrac{1}{2\cdot\xi} = \sigma^2, \tag{B.3}$$

i.e., the variance.

The information in (B.1) and (B.2) completely defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}_{<0}$ and the domain of possible parameters is

$\operatorname{int}\mathscr{T} = \mathbb{R}_{>0}$. The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and t an element of $\operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\int_{\Xi}\left(\sqrt{-2\cdot\xi}\right)^{s}\cdot\exp^{s}\langle\xi\,|\,t\rangle\;\mathrm{d}\xi.$$

It can be calculated more easily when starting from the noncanonical parameterization given by the function $\psi' \coloneqq \lambda : \mathbb{R}_{>0} ; -\frac{1}{2}\cdot\lambda$, where λ is the centered normal distribution's precision (related to the standard deviation σ by $\lambda = 1/\sigma^2$):

$$\begin{aligned}\mathsf{c}(s,t) &= 1/\int_{\mathbb{R}_{>0}}\lambda^{\frac{s}{2}}\cdot\exp^{s}\left(-\tfrac{1}{2}\cdot\lambda\cdot t\right)\cdot\left|\left|(\nabla\psi')\lambda\right|\right|\mathrm{d}\lambda\\ &= 1/\int_{\mathbb{R}_{>0}}\lambda^{\frac{s}{2}}\cdot\exp\left(-\tfrac{s}{2}\cdot t\cdot\lambda\right)\cdot\tfrac{1}{2}\,\mathrm{d}\lambda\\ &= 2\cdot\frac{\left(\frac{s}{2}\cdot t\right)^{\frac{s+3}{2}}}{\Gamma\frac{s+3}{2}},\end{aligned}\tag{B.4}$$

which is easily found after recognizing that the integral is proportional to one over the normalization factor of $\mathrm{Ga}\left(\bullet \;\middle|\; \frac{s+3}{2}, \frac{s}{2}\cdot t\right)$ (cf. $(4.31)_{169}$). So the prior parametric linear prevision is just this gamma prevision.

Using $(4.26)_{167}$, (B.4), and $(B.2)_{\frown}$, we can obtain the immediate prior predictive prevision. Its probability density for z in $\mathbb{R}$ is

$$\frac{\frac{1}{\sqrt{2\cdot\pi}}\cdot 2\cdot\frac{\left(\frac{s}{2}\cdot t\right)^{\frac{s+3}{2}}}{\Gamma\frac{s+3}{2}}}{2\cdot\frac{\left(\frac{s+1}{2}\cdot\left(\frac{s\cdot t+z^2}{s+1}\right)\right)^{\frac{s+1+3}{2}}}{\Gamma\frac{s+1+3}{2}}} = \frac{1}{\sqrt{2\cdot\pi}}\cdot\frac{\Gamma\frac{s+3+1}{2}}{\Gamma\frac{s+3}{2}}\cdot\frac{\left(\frac{s}{2}\cdot t\right)^{\frac{s+3+1}{2}}\cdot\left(\frac{s}{2}\cdot t\right)^{-\frac{1}{2}}}{\left(\frac{s+1}{2}\cdot\left(\frac{s\cdot t+z^2}{s+1}\right)\right)^{\frac{s+3+1}{2}}}$$

after some tedious manipulations this can be rewritten as

$$= \frac{1}{\sqrt{s\cdot t}}\cdot\frac{\Gamma\frac{s+3+1}{2}}{\Gamma\frac{s+3}{2}\cdot\sqrt{\pi}}\cdot\left(1+\frac{1}{s\cdot t}\cdot z^2\right)^{-\frac{s+3+1}{2}},$$

which is the probability density of the centered Student's distribution with prevision $\mathrm{St}\left(\bullet \;\middle|\; s+3, 0, \sqrt{t}\cdot\sqrt{\frac{s}{s+3}}\right)$. The posterior is obtained by the typical substitution (cf. §4.1.6$_{168}$).

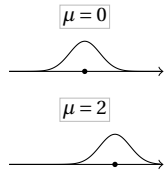
Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for centered normal sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \mathbb{R}_{>0}$. As this set is already bounded from below, the simplest choice is to only take an upper bound for t, i.e., an upper bound on the variance.

B.1.2 *Scaled normal sampling*

The scaled normal family is another subset of the normal family we saw in §4.1.2$_{159}$, now consisting of the elements for which the standard deviation is 1. It can be used whenever the standard deviation $\sigma : \mathbb{R}$ is

known; one just needs to do a coordinate scaling with a factor $1/\sigma$. Each member has $\mathscr{X} := \mathbb{R}$ as a sample space and is parameterized by its mean, i.e., a $\phi := \mu$ in $\Phi := \mathbb{R}$. The scaled normal linear prevision is $\mathrm{Nl}(\circledcirc|\mu, 1)$ (cf. $(4.9)_{159}$).

Two scaled normal densities (plot restricted to $[-5,5]$): (the mean μ is indicated with a dot)



Following the derivation in §4.1.2$_{159}$, we can easily derive the likelihood in exponential family form: let z be an observed real sample, then this likelihood can be written as

$$\begin{aligned}\tfrac{1}{\sqrt{2\cdot\pi}}\cdot\exp\bigl(-\tfrac{1}{2}\cdot(z-\mu)^2\bigr) &= \tfrac{1}{\sqrt{2\cdot\pi}}\cdot\exp\bigl(-\tfrac{z^2}{2}\bigr)\cdot\exp(z\cdot\mu)\cdot\exp\bigl(-\tfrac{\mu^2}{2}\bigr)\\ &= \exp\bigl(-\tfrac{\mu^2}{2}\bigr)\cdot\exp\langle\mu|z\rangle\cdot\tfrac{\exp\bigl(-\frac{z^2}{2}\bigr)}{\sqrt{2\cdot\pi}}.\end{aligned}$$

Comparing this last expression with $(4.1)_{156}$, we see that the Euclidean vector space for this family has dimension $d := 1$ and that

$$\begin{aligned}\tau &:= \mathrm{id}_{\mathbb{R}}, \quad \text{so} \quad \mathscr{T} := \mathrm{co}\,\mathbb{R} = \mathbb{R},\\ \psi &:= \mathrm{id}_{\mathbb{R}}, \quad \text{so} \quad \Xi := \mathbb{R},\end{aligned} \tag{B.5}$$

(so this is a linear canonical exponential family) and also that

$$\mathsf{a} := z:\mathbb{R}\,;\,\tfrac{\exp\bigl(-\frac{z^2}{2}\bigr)}{\sqrt{2\cdot\pi}} \quad \text{and} \quad \mathsf{b} := \xi:\mathbb{R}\,;\,\exp\bigl(-\tfrac{\xi^2}{2}\bigr). \tag{B.6}$$

The cumulant function for the scaled normal family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi:\mathbb{R}\,;\,\tfrac{\xi^2}{2}.$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find – letting $\xi := \psi\mu$ – that

$$\mathrm{Nl}(\tau|\mu,1) = (\nabla\kappa)\xi = \xi = \mu, \tag{B.7}$$

i.e., the mean.

The information in (B.5) and (B.6) completely defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}$ and the domain of possible parameters is $\mathrm{int}\,\mathscr{T} = \mathbb{R}$. The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and t an element of $\mathrm{int}\,\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\Xi}\exp^s\bigl(-\frac{\xi^2}{2}\bigr)\cdot\exp^s\langle\xi|t\rangle\,\mathrm{d}\xi$$

Because of its linear canonicality, we do not need to change parameterizations to find this normalization factor:

$$\begin{aligned}&= 1/\textstyle\int_{\Xi}\exp\bigl(-\frac{s}{2}\cdot(\xi^2-2\cdot\xi\cdot t)\bigr)\,\mathrm{d}\xi\\ &= \sqrt{\tfrac{s}{2\cdot\pi}}\cdot\exp\bigl(-\tfrac{s}{2}\cdot t^2\bigr),\end{aligned} \tag{B.8}$$

which is easily found after completing the square and recognizing that the integral is proportional to one over the normalization factor of the

normal prevision $\mathrm{Nl}\big(\bullet \,\big|\, t, \frac{1}{\sqrt{s}}\big)$ (cf. $(4.9)_{159}$). So the prior parametric linear prevision is just this normal prevision.

Using $(4.26)_{167}$, $(B.8)_{\frown}$, and $(B.6)_{\frown}$, we can obtain the immediate prior predictive prevision. Its probability density for z in $\mathbb{R}$ is

$$\frac{\frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\cdot\pi}}\cdot\sqrt{\frac{s}{2\cdot\pi}}\cdot\exp\left(-\frac{s}{2}\cdot t^2\right)}{\sqrt{\frac{s+1}{2\cdot\pi}}\cdot\exp\left(-\frac{s+1}{2}\cdot\left(\frac{s\cdot t+z}{s+1}\right)^2\right)} = \frac{1}{\sqrt{2\cdot\pi}\cdot\sqrt{\frac{s+1}{s}}}\cdot\frac{\exp\left(-\frac{1}{2}\cdot(s\cdot t^2+z^2)\right)}{\exp\left(-\frac{1}{2}\cdot\left(\frac{s^2}{s+1}\cdot t^2-2\cdot\frac{s}{s+1}\cdot z\cdot t+\frac{1}{s+1}\cdot z^2\right)\right)}$$

after some manipulations this can be rewritten as

$$= \frac{1}{\sqrt{2\cdot\pi}\cdot\sqrt{1+\frac{1}{s}}}\cdot\exp\left(-\frac{(z-t)^2}{2\cdot(1+\frac{1}{s})}\right),$$

which is the probability density of a normal distribution with prevision $\mathrm{Nl}\big(\bullet \,\big|\, t, \sqrt{1+1/s}\big)$. The posterior is obtained by the typical substitution (cf. $\S4.1.6_{168}$).

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for scaled normal sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \mathbb{R}$. The simplest choice is to only take an upper bound for $|t|$, i.e., an upper bound on the absolute value of the mean.

B.1.3 *Multivariate normal sampling*

The family of ℓ-variate $(\ell : \mathbb{N}_{>0})$ normal distributions [Kotz et al. 2000, $\S45_{105}$] is the last of the normal families we are going to treat. Each member has $\mathscr{X} := \mathbb{R}^\ell$ as a sample space and is parameterized by its mean vector and precision matrix – the inverse of the covariance matrix –, i.e., a $\phi := (\mu, \Lambda)$ in $\Phi := \mathbb{R}^\ell \times \mathbb{R}^{\ell\times\ell}_{\mathrm{pod}}$, where $\mathbb{R}^{\ell\times\ell}$ is the set of square ℓ-dimensional matrices and the symmetric positive definiteness predicate pod is defined by (let M be a square matrix of size ℓ)

$$\operatorname{pod} M \Leftrightarrow M = M^\mathsf{T} \wedge \forall v : (\mathbb{R}^\ell)_{\neq 0} ; v^\mathsf{T} M v > 0. \tag{B.9}$$

Recall that any positive definite matrix is invertible. We later also need the related concept of symmetric positive semidefiniteness and symmetric negative definiteness, expressed by the predicates psd and ned respectively, which are defined by

$$\operatorname{psd} M \Leftrightarrow M = M^\mathsf{T} \wedge \forall v : (\mathbb{R}^\ell)_{\neq 0} ; v^\mathsf{T} M v \geq 0, \tag{B.10}$$

$$\operatorname{ned} M \Leftrightarrow M = M^\mathsf{T} \wedge \forall v : (\mathbb{R}^\ell)_{\neq 0} ; v^\mathsf{T} M v < 0. \tag{B.11}$$

Let f be any measurable gamble on $\mathbb{R}^\ell$; the multivariate normal linear prevision is then defined by

$$\mathrm{Nr}(f|\mu,\Lambda) := \int_{\mathscr{X}} fz \cdot \sqrt{\tfrac{|\Lambda|}{(2\cdot\pi)^\ell}} \cdot \exp\bigl(-\tfrac{1}{2}\cdot(z-\mu)^\mathsf{T}\Lambda(z-\mu)\bigr)\,\mathrm{d}z. \tag{B.12}$$

We start by rewriting the corresponding likelihood function in exponential family form: let $z:\mathbb{R}^\ell$ be an observed sample, then this likelihood can be written as

The trace function tr applied to a square matrix returns the sum of its main diagonal's components.

$$\begin{aligned}\sqrt{\tfrac{|\Lambda|}{(2\cdot\pi)^\ell}} \cdot \exp\bigl(-\tfrac{1}{2}\cdot(z-\mu)^\mathsf{T}\Lambda(z-\mu)\bigr) &= \sqrt{\tfrac{|\Lambda|}{(2\cdot\pi)^\ell}} \cdot \exp\bigl(-\tfrac{\mathrm{tr}(\Lambda zz^\mathsf{T})}{2} + \tfrac{\mu^\mathsf{T}\Lambda z}{\sigma^2} - \tfrac{\mu^\mathsf{T}\Lambda\mu}{2}\bigr)\\ &= \exp\bigl(-\tfrac{\mu^\mathsf{T}\Lambda\mu}{2} + \ln|\Lambda|\bigr)\cdot\exp\bigl\langle\mu^\mathsf{T}\Lambda, -\tfrac{\Lambda}{2}\,\big|\, z, zz^\mathsf{T}\bigr\rangle\cdot\tfrac{1}{\sqrt{2\cdot\pi}^\ell}.\end{aligned}$$

Here, we used the fact that $v^\mathsf{T}Mv = \mathrm{tr}(Mvv^\mathsf{T})$ for square matrices M and real vectors v of the same dimension [Bernstein 2005, $(2.2.26)_{22}$]. Also, in contrast to all the other scalar products we have used for exponential family notation above or will use below, it is not just a product or sum of products: the second part of the scalar product involves taking the trace of a matrix product (cf. Frobenius norm [see, e.g., Bernstein 2005, $\S 9.2.1_{347}$]). This expression and the ensuing derivations below can be rewritten in the vectorial language, however, but we opted not to do this to stress the strong parallels between the univariate and multivariate normal families.

Comparing the expression with $(4.1)_{156}$ and taking the symmetries of Λ and the positive semidefinite dyadic (or outer) product zz^T, we see that the Euclidean vector space for this family has $d := \ell + \frac{\ell\cdot(\ell+1)}{2} = \frac{\ell\cdot(\ell+3)}{2}$ as its dimension and that

$$\begin{aligned}&\tau := z:\mathbb{R}^\ell;(z,zz^\mathsf{T}), \quad \text{so} \quad \mathscr{T} := \mathrm{co}\{z, zz^\mathsf{T} \mid z:\mathbb{R}^\ell\}\\ &\qquad\qquad = \bigl\{t:\mathbb{R}^\ell\times\mathbb{R}^{\ell\times\ell}_{\mathrm{psd}} \,\big|\, t_2 - t_1 t_1^\mathsf{T} \in \mathbb{R}^{\ell\times\ell}_{\mathrm{psd}}\bigr\},\\ &\psi := (\mu,\Lambda):\mathbb{R}^\ell\times\mathbb{R}^{\ell\times\ell}_{\mathrm{pod}};(\Lambda\mu, -\tfrac{\Lambda}{2}), \quad \text{so} \quad \Xi := \mathbb{R}^\ell\times\mathbb{R}^{\ell\times\ell}_{\mathrm{ned}},\end{aligned} \tag{B.13}$$

Some extra clarification of the definition of the set $\mathscr{T}$ of pseudomeans, each of which has one matrix component, is in order:

(i) Any symmetric positive semidefinite matrix can be seen to be a convex combination of dyadic products via its spectral decomposition [Bernstein 2005, $\S 5.12.10_{205}$ and $\S 8.1_{263}$].

(ii) By construction, for any element t of $\mathscr{T}$, the difference $t_2 - t_1 t_1^\mathsf{T}$ is a covariance matrix (i.e., of the probability measure described by the convex combination determining t) and thus a symmetric positive semidefinite matrix, as these two concepts coincide [see, e.g., Wikipedia 2008*b*].

From the expression of the likelihood in exponential-family form, we

can also deduce that

$$\mathsf{a} := \mathbb{R}\,;\, \tfrac{1}{\sqrt{2\cdot\pi}^{\ell}} \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R}^{\ell} \times \mathbb{R}^{\ell\times\ell}_{\text{ned}}\,;\, \exp\Big(\tfrac{1}{2}\cdot\Big(\tfrac{\xi_1^{\mathsf{T}}\xi_2^{-1}\xi_1}{2} + \ln\big(-2\cdot|\xi_2|\big)\Big)\Big). \tag{B.14}$$

The cumulant function for the multivariate normal family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi : \mathbb{R}^{\ell} \times \mathbb{R}^{\ell\times\ell}_{\text{ned}}\,;\, -\tfrac{1}{2}\cdot\big(\tfrac{\xi_1^{\mathsf{T}}\xi_2^{-1}\xi_1^{\mathsf{T}}}{2} + \ln(-2\cdot|\xi_2|)\big).$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find – letting $\xi := \psi(\mu,\Lambda)$ – that

$$\mathrm{Nr}(\tau|\mu,\Lambda) = (\nabla\kappa)\xi \tag{B.15}$$

$$= -\tfrac{1}{2}\cdot\big(\xi_2^{-1}\xi_1, -\tfrac{\xi_2^{-1}\xi_1\xi_1^{\mathsf{T}}\xi_2^{-1}}{2} + \xi_2^{-1}\big) = (\mu, \mu\mu^{\mathsf{T}} + \Lambda^{-1}), \tag{B.16}$$

i.e., a vector consisting of the mean (vector) and the (matrix of) second-order noncentral moments.

The information in $(\text{B.13})_{\frown}$ and (B.14) completely defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}^{\ell} \times \mathbb{R}^{\ell\times\ell}_{\text{ned}}$ and the domain of possible parameters is

$$\operatorname{int}\mathscr{T} = \big\{t : \mathbb{R}^{\ell} \times \mathbb{R}^{\ell\times\ell}_{\text{pod}} \;\big|\; t_2 - t_1\,t_1^{\mathsf{T}} \in \mathbb{R}^{\ell\times\ell}_{\text{pod}}\big\}.$$

Every neighborhood (Frobenius norm) of a matrix in $\mathbb{R}^{\ell\times\ell}_{\text{psd}\wedge\neg\text{pod}}$ contains elements of $\mathbb{R}^{\ell\times\ell}_{\neg\text{psd}}$: the zero eigenvalue is arbitrarily close to the negative reals.

The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and $t : \operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\Xi} \exp^{s}\Big(\tfrac{1}{2}\cdot\Big(\tfrac{\xi_1^{\mathsf{T}}\xi_2^{-1}\xi_1}{2} + \ln\big(-2\cdot|\xi_2|\big)\Big)\Big)\cdot\exp^{s}\langle\xi|t\rangle\,\mathrm{d}\xi.$$

It can be calculated more easily when starting from the original non-canonical parameterization ψ:

$$\begin{aligned}\mathsf{c}(s,t) &= 1/\textstyle\int_{\mathbb{R}^{\ell}\times\mathbb{R}^{\ell\times\ell}_{\text{pod}}} \exp^{s}\Big(\tfrac{1}{2}\cdot(-\mu^{\mathsf{T}}\Lambda\mu + \ln|\Lambda|) + t_1^{\mathsf{T}}\Lambda\mu + \mathrm{tr}\big(-\tfrac{\Lambda}{2}\,t_2\big)\Big)\\ &\qquad\qquad \cdot\big|(\nabla\psi')(\mu,\Lambda)\big|\,\mathrm{d}\mu\,\mathrm{d}\Lambda\\ &= 1/\textstyle\int_{\mathbb{R}^{\ell\times\ell}_{\text{pod}}}\Big(\int_{\mathbb{R}^{\ell}} |\Lambda|^{\frac{s}{2}}\cdot\exp^{s}\Big(-\tfrac{1}{2}\cdot(\mu^{\mathsf{T}}\Lambda\mu - t_1^{\mathsf{T}}\Lambda\mu - \mu^{\mathsf{T}}\Lambda t_1) - \mathrm{tr}\big(\tfrac{\Lambda}{2}\,t_2\big)\Big)\\ &\qquad\qquad \cdot\Big|\begin{matrix}\Lambda & \mathbb{O}\\ \mu^{\mathsf{T}} & -\frac{1}{2}\end{matrix}\Big|\,\mathrm{d}\mu\Big)\,\mathrm{d}\Lambda;\end{aligned}$$

completing the square of the exponent's first term and splitting the exponent gives

$$\begin{aligned}&= 1/\textstyle\int_{\mathbb{R}^{\ell\times\ell}_{\text{pod}}}\Big(\int_{\mathbb{R}^{\ell}} |\Lambda|^{\frac{s}{2}}\cdot\exp^{s}\big(-\tfrac{1}{2}\cdot(\mu - t_1)^{\mathsf{T}}\Lambda(\mu - t_1)\big)\\ &\qquad\qquad \cdot\exp^{s}\big(-\mathrm{tr}\big(\tfrac{1}{2}\cdot(t_2 - t_1\,t_1^{\mathsf{T}})\Lambda\big)\cdot\tfrac{1}{2}\cdot|\Lambda|\,\mathrm{d}\mu\Big)\,\mathrm{d}\Lambda\end{aligned}$$

$$= 2/\int_{\mathbb{R}^{\ell\times\ell}_{\text{pod}}}\Big(\int_{\mathbb{R}^\ell}|\Lambda|^{\frac{1}{2}}\cdot\exp\big(-\tfrac{s}{2}\cdot(\mu-t_1)^\intercal\Lambda(\mu-t_1)\big)\,\mathrm{d}\mu\Big)$$
$$\cdot|\Lambda|^{\frac{s+1}{2}}\cdot\exp\big(-\operatorname{tr}\big(\tfrac{s}{2}\cdot(t_2-t_1t_1^\intercal)\Lambda\big)\,\mathrm{d}\Lambda$$
$$= 2\cdot\frac{\big(\frac{s}{2}\cdot|t_2-t_1t_1^\intercal|\big)^{\frac{s+\ell+2}{2}}}{\Gamma_\ell\,\frac{s+\ell+2}{2}}\cdot\sqrt{\tfrac{s}{(2\cdot\pi)^\ell}}. \tag{B.17}$$

The last step follows from the fact that the inner integral is proportional to $\mathrm{Nr}(\mathbb{R}^\ell|t_1,s\cdot\Lambda)$ and therefore the entire double integral is proportional to $\mathrm{Wi}\big(\mathbb{R}^{\ell\times\ell}_{\text{pod}}\,\big|\,\frac{s+\ell+2}{2},\frac{s}{2}\cdot(t_2-t_1t_1^\intercal)\big)$, where (let $f:\breve{\mathscr{L}}_{\mathbb{R}^{\ell\times\ell}_{\text{pod}}}$

$$\mathrm{Wi}(f|\alpha,\beta)\coloneqq\int_{\mathbb{R}^{\ell\times\ell}_{\text{pod}}}f r\cdot\tfrac{|\beta|^\alpha}{\Gamma_\ell\alpha}\cdot|r|^{\alpha-\frac{\ell+1}{2}}\cdot\exp\big(-\operatorname{tr}(\beta r)\big)\,\mathrm{d}r \tag{B.18}$$

is the Wishart linear prevision [see, e.g., Bernardo & Smith 1994, §3.2.5$_{138}$], a multivariate generalization of the gamma prevision, with parameters $\alpha:\mathbb{R}_{>(\ell-1)/2}$ and $\beta:\mathbb{R}^{\ell\times\ell}_{\text{pod}}$.

The generalized gamma function $\Gamma_\bullet$: (let $\ell:\mathbb{N}_{>0}$ and $\alpha:\mathbb{R}_{>(\ell-1)/2}$)

$$\Gamma_\ell\alpha\coloneqq\pi^{\frac{\ell\cdot(\ell-1)}{4}}\cdot\textstyle\prod_{i:1..\ell}\Gamma\frac{2\cdot\alpha+1-i}{2}.$$

For obvious reasons the conjugate family is called the normal-Wishart family. The normal-Wishart linear prevision [see, e.g., Bernardo & Smith 1994, §3.2.5$_{140}$] is defined by (now let f be a measurable gamble on the space $\mathbb{R}^\ell\times\mathbb{R}^{\ell\times\ell}_{\text{pod}}$)

$$\mathrm{Nw}(f|\mu,\rho,\alpha,\beta)\coloneqq\mathrm{Wi}\big(\mathrm{Nr}(f|\mu,\rho\cdot\bullet)\,\big|\,\alpha,\beta\big), \tag{B.19}$$

with parameters restricted as before: μ – not to be confused with earlier uses of this symbol – can be any vector in $\mathbb{R}^\ell$ and ρ must be a strictly positive real number. So the prior parametric linear prevision is

$$\mathrm{Nw}\big(\bullet\,\big|\,t_1,s,\tfrac{s+\ell+2}{2},\tfrac{s}{2}\cdot(t_2-t_1t_1^\intercal)\big). \tag{B.20}$$

(Note that not every normal-Wishart prevision belongs to the conjugate family.) The posterior is obtained by the typical substitution (cf. §4.1.6$_{168}$).

We are not going to derive an expression for any predictive prevision here, as the per-subsection critical tedious manipulation limit would then surely be crossed.

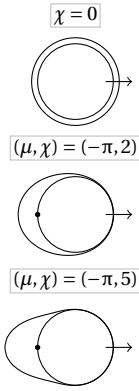
All it takes to define imprecise-probabilistic inference models of the type of (4.40)–(4.43)$_{174}$ for multivariate normal sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of

$$\operatorname{int}\mathscr{T}=\big\{t:\mathbb{R}^\ell\times\mathbb{R}^{\ell\times\ell}_{\text{pod}}\,\big|\,t_2-t_1t_1^\intercal\in\mathbb{R}^{\ell\times\ell}_{\text{pod}}\big\}.$$

If sufficient information is available, specific bounds may be imposed for each of the components of the means vector (first part of t) and the components of the matrix of noncentral second moments (second part of t). In case almost no information is available, specifying a very high upper bound on the Frobenius norm of the matrix of noncentral second moments is a compact way to obtain a model for near-ignorance.

B.1.4 *Von Mises planar directional sampling*

Three von Mises densities: (the density is represented by the distance to the circle, the mean angle μ is indicated with a dot)



The von Mises family of distributions is typically used to model directional sampling in the plane [Fisher et al. 1987, §4.3.2$_{81}$]; each member has an interval of length $2\cdot\pi$ as a sample space and is parameterized by its mean angle (measured counterclockwise from the reference direction $(0,1)$) and so-called concentration, i.e., a $\phi := (\mu,\chi)$ in $\Phi := \,]-\pi,\pi] \times \mathbb{R}_{\geq 0}$ (the exact choice of interval for μ is of course arbitrary). Let f be any measurable $2\cdot\pi$-periodic gamble on $\mathbb{R}$, the von Mises linear prevision is then defined by

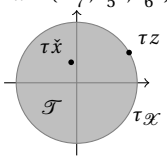
$$\mathrm{vM}(f|\mu,\chi) := \int_{]\mu-\pi,\mu+\pi]} f z \cdot \tfrac{1}{2\cdot\pi\cdot I_0\chi} \cdot \exp\big(\chi\cdot\cos(z-\mu)\big)\,\mathrm{d}z, \tag{B.21}$$

where I_0 is the zeroth-order modified Bessel function [Abramowitz & Stegun 1972, §9.6$_{374}$], with $I_0 0 = 1$. When $\chi = 0$, this prevision reduces to the uniform one, which has no mean angle. The integration bounds were chosen in a way to limit some technical issues related to the choice of sample space; it results in an only seemingly μ-dependent sample space $\mathscr{X} := \,]\mu-\pi,\mu+\pi]$.

We start by rewriting the corresponding likelihood function in exponential family form: let z be an observed angle, then this likelihood can be written as

$$\begin{aligned}\tfrac{1}{2\cdot\pi\cdot I_0\chi} \cdot \exp\big(\chi\cdot\cos(z-\mu)\big) &= \tfrac{1}{2\cdot\pi\cdot I_0\chi} \cdot \exp(\chi\cdot\cos\mu\cdot\cos z + \chi\cdot\sin\mu\cdot\sin z)\\ &= \tfrac{1}{I_0\chi} \cdot \exp\langle \chi\cdot\cos\mu, \chi\cdot\sin\mu | \cos z, \sin z\rangle \cdot \tfrac{1}{2\cdot\pi}.\end{aligned}$$

Take $z := \frac{\pi}{6}$ and $\check{x} := (-\frac{\pi}{7}, \frac{3\cdot\pi}{5}, \frac{5\cdot\pi}{6})$:



Comparing this last expression with $(4.1)_{156}$, we see that the Euclidean vector space for this family has dimension $d := 2$ and that (let $\mathbb{S} = \{\tau_{\mathscr{X}}\}$ be the unit circle and $\|\bullet\| := \sqrt{\sum \bullet^2}$ the Euclidean norm)

$$\begin{aligned}&\tau := (\cos,\sin), \quad \text{so} \quad \mathscr{T} := \mathrm{co}\,\mathbb{S} = \big\{t : \mathbb{R}^2 \,\big|\, \|t\| \leq 1\big\},\\ &\psi := (\mu,\chi) : \,]-\pi,\pi] \times \mathbb{R}_{\geq 0} ; (\chi\cdot\cos\mu, \chi\cdot\sin\mu), \quad \text{so} \quad \Xi := \mathbb{R}^2,\end{aligned} \tag{B.22}$$

and also that

$$\mathsf{a} := \mathscr{X} ; \tfrac{1}{2\cdot\pi} \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R}^2 ; \tfrac{1}{I_0\|\xi\|}. \tag{B.23}$$

The cumulant function for the von Mises family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi : \mathbb{R}^2 ; \ln\big(I_0\|\xi\|\big).$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find – letting $\xi := \psi(\mu,\chi)$ – that

$$\mathrm{vM}(\tau|\mu,\chi) = (\nabla\kappa)\xi = \tfrac{I_1\|\xi\|}{I_0\|\xi\|} \cdot \tfrac{1}{2} \cdot \tfrac{2\cdot\xi}{\|\xi\|} = \tfrac{I_1\chi}{I_0\chi} \cdot (\cos\mu, \sin\mu), \tag{B.24}$$

where we have used the fact that $\mathrm{D}I_0 = I_1$ [Abramowitz & Stegun 1972, §9.6.27$_{376}$].

The information in (B.22) and (B.23) completely defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}^2$ and the parameter set is $\mathrm{int}\,\mathscr{T} = \big\{t : \mathbb{R}^2 \,\big|\, \|t\| < 1\big\}$.

The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and $t : \operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\Xi} \frac{1}{(I_0\|\xi\|)^s} \cdot \exp^s\langle\xi|t\rangle \,\mathrm{d}\xi.$$

It can be calculated more easily when starting from the original non-canonical parameterization ψ:

$$\begin{aligned}\mathsf{c}(s,t) &= 1/\textstyle\int_{]-\pi,\pi]\times\mathbb{R}_{\geq 0}} \frac{1}{(I_0\chi)^s} \cdot \exp\big(s\cdot\sum(\chi\cdot\cos\mu,\chi\cdot\sin\mu)\cdot t\big) \\ &\qquad\qquad \cdot\big|\big|(\nabla\psi)(\mu,\chi)\big|\big| \,\mathrm{d}\mu\,\mathrm{d}\chi \\ &= 1/\textstyle\int_{\mathbb{R}_{\geq 0}} \frac{1}{(I_0\chi)^s} \cdot \Big(\int_{-\pi}^{\pi} \exp\langle s\cdot t\cdot\chi|\cos\mu,\sin\mu\rangle \cdot \Big|\Big|\begin{smallmatrix}\cos\mu & \sin\mu \\ \chi\cdot\sin\mu & -\chi\cdot\cos\mu\end{smallmatrix}\Big|\Big| \,\mathrm{d}\mu\Big) \,\mathrm{d}\chi \\ &= 1/\textstyle\int_{\mathbb{R}_{\geq 0}} \chi \cdot \frac{1}{(I_0\chi)^s} \cdot \Big(\int_{\measuredangle t-\pi}^{\measuredangle t+\pi} \exp\langle s\cdot t\cdot\chi|\cos\mu,\sin\mu\rangle \,\mathrm{d}\mu\Big) \,\mathrm{d}\chi \\ &= 1/\textstyle\int_{\mathbb{R}_{\geq 0}} \chi \cdot \frac{2\cdot\pi\cdot I_0(s\cdot\|t\|\cdot\chi)}{(I_0\chi)^s} \,\mathrm{d}\chi, \end{aligned} \tag{B.25}$$

The angle function $\measuredangle$ returns the angle of its argument (an element of $(\mathbb{R}^2)_{\neq 0}$) with relation to the reference direction $(0,1)$.

where the last step consisted in recognizing that the inner integral is one over the normalization factor of $\mathrm{vM}\big(\bullet \,\big|\, \measuredangle t, s\cdot\|t\|\cdot\chi\big)$. Further simplification of the normalization factor is not evident.

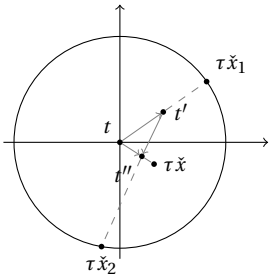
So the prior parametric linear prevision is

$$\mathrm{vC}(\bullet|s,t) :\propto \textstyle\int_{\mathbb{R}_{\geq 0}} \mathrm{vM}\big(\bullet \,\big|\, \measuredangle t, s\cdot\|t\|\cdot\chi\big)\cdot\chi\cdot\frac{I_0(s\cdot\|t\|\cdot\chi)}{(I_0\chi)^s}\,\mathrm{d}\chi. \tag{B.26}$$

Considering that we do not have the normalization factor (B.25) available in closed form, we are not going to give explicit expressions for any predictive prevision.

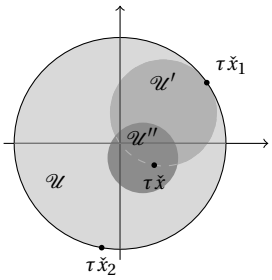
On the side, we have given a graphical illustration of how the parameters are updated. We use a number of pseudocounts $s := 1$ and an observed sample $\check{x}$ of length 2. Thus, taking into account the definition of τ, the two updated mean single-sample sufficient statistic parameters are

$$\begin{aligned} t' &:= \tfrac{1}{2}\cdot t + \tfrac{1}{2}\cdot(\cos\check{x}_1, \sin\check{x}_1), \\ t'' &:= \tfrac{2}{3}\cdot t' + \tfrac{1}{3}\cdot(\cos\check{x}_2, \sin\check{x}_2) \\ &= \tfrac{1}{3}\cdot t + \tfrac{2}{3}\cdot(\cos\check{x}_1 + \cos\check{x}_2, \sin\check{x}_1 + \sin\check{x}_2). \end{aligned}$$



Note how the proportions appearing in these expressions are reflected in the illustration.

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for von Mises sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \big\{t : \mathbb{R}^2 \,\big|\, \|t\| < 1\big\}$. As this is already a bounded set, the easiest choice is to take $\mathscr{U} := \operatorname{int}\mathscr{T}$ and start with a near-ignorance prior, unless there is so much prior information about the von Mises process that would make it worthwhile to specify bounds.



On the side, we have given a graphical illustration of how the parameters of the imprecise-probabilistic model are updated. We again use a number of pseudocounts $s := 1$, the initial set of mean single-sample sufficient statistics is $\mathscr{U} := \operatorname{int}\mathscr{T}$, which corresponds to a near-ignorance prior, and take an observed sample $\check{x}$ of length 2. Thus, taking into account the definition of τ, the two updated sets of mean single-sample sufficient statistics are

$$\begin{aligned}\mathscr{U}' &:= \tfrac{1}{2}\cdot\mathscr{U} + \tfrac{1}{2}\cdot(\cos\check{x}_1, \sin\check{x}_1),\\ \mathscr{U}'' &:= \tfrac{2}{3}\cdot\mathscr{U}' + \tfrac{1}{3}\cdot(\cos\check{x}_2, \sin\check{x}_2)\\ &= \tfrac{1}{3}\cdot\mathscr{U} + \tfrac{2}{3}\cdot(\cos\check{x}_1 + \cos\check{x}_2, \sin\check{x}_1 + \sin\check{x}_2).\end{aligned}$$

Note again how the proportions appearing in these expressions are reflected in the illustration.

Considering $\tau = (\cos, \sin)$, the set of gambles for which we know the predictive lower and upper prevision can be efficiently calculated (cf. $(4.46)_{178}$) consists more or less all the sinusoidal functions with angular frequency 1 and arbitrary phase. After a cursory look, this does not seem a very interesting set of gambles. Similarly, as

$$(\nabla\kappa)\big(\psi(\mu,\chi)\big) = \tfrac{I_1\chi}{I_0\chi}\cdot(\cos\mu, \sin\mu),$$

the same holds for the set of gambles for which we know the parametric lower and upper prevision can be efficiently calculated.

This does not mean that calculating the lower and upper prevision of all other gambles is hard. For example, the lower and upper parametric prevision of the mean: let $g := \mathrm{id}_{]-\pi,\pi]}\cdot I^{\mathbb{R}_{\geq 0}}$ and use the parameters and observations of the graphical illustration above, then

$$\begin{aligned}[\underline{\mathrm{vC}}(g|1,\mathscr{U}), \overline{\mathrm{vC}}(g|1,\mathscr{U})] &= [-\pi,\pi],\\ [\underline{\mathrm{vC}}(g|2,\mathscr{U}'), \overline{\mathrm{vC}}(g|2,\mathscr{U}')] &= [-\tfrac{11}{36}\cdot\pi, \tfrac{25}{36}\cdot\pi],\\ [\underline{\mathrm{vC}}(g|3,\mathscr{U}''), \overline{\mathrm{vC}}(g|3,\mathscr{U}'')] &= [-\pi,\pi],\end{aligned}$$

where $\underline{\mathrm{vC}}(\bullet|s,\mathscr{U}) := \inf_{t:\mathscr{U}} \mathrm{vC}(\bullet|s,t)$. Remark how the inferences for this gamble dilate after the second observation.

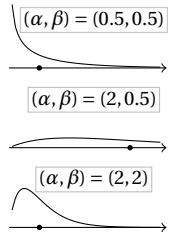
❧

The von Mises family (also called the circular normal family) is part of a class of families – one for every dimensionality – used for modeling directional sampling, the von Mises–Fisher class [see, e.g., Dhillon & Sra 2003]. Apart from the von Mises family, they are usually written in terms of unit vectors and not angles; Barndorff-Nielsen [1978, Ex. 8.1_{113}] shows that as such they are all linear regular exponential families. Mardia & El-Atoum [1976] give a short discussion of Bayesian inference for these families.

B.1.5 *Gamma sampling*

The gamma family is commonly used, for example when modeling waiting times [Johnson & Kotz 1970, §17.3$_{171}$]. Each member $\mathrm{Ga}(\bullet|\alpha,\beta)$ (cf. (4.31)$_{169}$) has $\mathscr{X} := \mathbb{R}_{>0}$ as a sample space and, as seen in §4.1.6$_{168}$, is parameterized by two strictly positive real parameters, a shape parameter α and a rate parameter β.

Some gamma densities (plot restricted to $]0,5]$): (the mean α/β is indicated with a dot)



Writing the likelihood in exponential family form is straightforward: let z be an observed waiting time, then the likelihood can be written as

$$\frac{\beta^\alpha}{\Gamma\alpha}\cdot z^{\alpha-1}\cdot\exp(-\beta\cdot z) = \frac{\beta^\alpha}{\Gamma\alpha}\cdot\exp\langle -\beta,\alpha-1|z,\ln z\rangle.$$

Comparing this last expression with (4.1)$_{156}$, we see that the Euclidean vector space for this family has dimension $d := 2$ and that

$$\begin{aligned}&\tau := z:\mathbb{R}_{>0}\,;(z,\ln z),\quad\text{so}\quad \mathscr{T} := \mathrm{co}\,\mathbb{R}^2_{>0} = \mathbb{R}^2_{>0},\\ &\psi := \alpha,\beta:\mathbb{R}^2_{>0}\,;(-\beta,\alpha-1),\quad\text{so}\quad \Xi := \mathbb{R}_{<0}\times\mathbb{R}_{>1},\end{aligned} \tag{B.27}$$

and that

$$\mathsf{a} := \mathbb{R}_{\geq 0}\,;1\quad\text{and}\quad \mathsf{b} := \xi:\mathbb{R}_{<0}\times\mathbb{R}_{>1}\,;\tfrac{(-\xi_1)^{\xi_2+1}}{\Gamma(\xi_2+1)}. \tag{B.28}$$

The cumulant function for the gamma family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi:\mathbb{R}_{<0}\times\mathbb{R}_{>1}\,;-(\xi_2+1)\cdot\ln(-\xi_1)+\ln\big(\Gamma(\xi_2+1)\big).$$

So, using (4.7)$_{159}$ and (4.8)$_{159}$, we find – letting $\xi := \psi(\alpha,\beta)$ – that

$$\mathrm{Ga}(\tau|\alpha,\beta) = (\nabla\kappa)\xi = (\tfrac{\alpha}{\beta},\Psi\alpha-\ln\beta), \tag{B.29}$$

The digamma function [Abramowitz & Stegun 1972, §6.4$_{260}$] is defined by $\Psi := \mathrm{D}(\ln\circ\Gamma)$.

a vector with as the first component the mean waiting time and as the second component a quantity that has no immediately unambiguous interpretation to me (perhaps "the mean logarithm of the waiting time", "the logarithm of the geometric mean waiting time", or even "the mean Neperian order of magnitude of the waiting time").

The information in (B.27) and (B.28) fully defines the canonical form (4.21)$_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}_{<0}\times\mathbb{R}_{>1}$ and the set of parameters is $\mathrm{int}\,\mathscr{T} = \mathbb{R}^2_{>0}$. The one thing we do not yet know explicitly is the normalization factor (4.22)$_{165}$: (let s be a strictly positive real number and $t:\mathrm{int}\,\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_\Xi\big(\frac{(-\xi_1)^{\xi_2+1}}{\Gamma(\xi_2+1)}\big)^s\cdot\exp^s\langle\xi|t\rangle\,\mathrm{d}\xi.$$

It can be calculated more easily when starting from the original non-canonical parameterization ψ:

$$\begin{aligned}&\mathsf{c}(s,t)\\ &\quad= 1/\textstyle\int_{\mathbb{R}^2_{>0}}\frac{(\beta^\alpha)^s}{(\Gamma\alpha)^s}\cdot\exp(-\beta\cdot s\cdot t_1+(\alpha-1)\cdot s\cdot t_2)\cdot\big|(\nabla\psi)(\alpha,\beta)\big|\,\mathrm{d}\alpha\,\mathrm{d}\beta\\ &\quad= 1/\textstyle\int_{\mathbb{R}_{>0}}\frac{\exp((\alpha-1)\cdot s\cdot t_2)}{(\Gamma\alpha)^s}\cdot\big(\int_{\mathbb{R}_{>0}}\beta^{\alpha\cdot s}\cdot\exp(-\beta\cdot s\cdot t_1)\,\mathrm{d}\beta\big)\cdot\left|\begin{matrix}0&1\\-1&0\end{matrix}\right|\,\mathrm{d}\alpha\end{aligned}$$

$$= 1/\int_{\mathbb{R}_{>0}} \frac{\Gamma(\alpha\cdot s+1)\cdot\exp((\alpha-1)\cdot s\cdot t_2)}{(\Gamma\alpha)^s\cdot(s\cdot t_1)^{\alpha\cdot s+1}}\,\mathrm{d}\alpha, \tag{B.30}$$

where the last step consisted in recognizing that the inner integral is one over the normalization factor of $\mathrm{Ga}(\circledast|\alpha\cdot s+1, s\cdot t_1)$. Further simplification of the normalization factor is not evident [also see Miller 1980].

So the prior parametric linear prevision is

$$\mathrm{Gc}(\circledast|s,t) :\propto \int_{\mathbb{R}_{>0}} \mathrm{Ga}(\circledast|\alpha\cdot s+1, s\cdot t_1)\cdot \frac{\Gamma(\alpha\cdot s+1)\cdot\exp((\alpha-1)\cdot s\cdot t_2)}{(\Gamma\alpha)^s\cdot(s\cdot t_1)^{\alpha\cdot s+1}}\,\mathrm{d}\alpha. \tag{B.31}$$

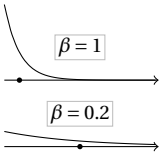
Given we do not have the normalization factor (B.30) available in closed form, we are not going to give explicit expressions for any predictive prevision.

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–(4.43)$_{174}$ for Gamma sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \mathbb{R}^2_{>0}$. As both components are already bounded from below, the simplest choice is to only take an upper bound for t, i.e., an upper bound on the mean waiting time and an upper bound on the mean order of magnitude of the waiting time.

The next family is a subfamily of the gamma family for which $\alpha := 1$; it presents fewer analytical difficulties.

B.1.6 *Exponential sampling*

Two exponential densities (plot restricted to $[0,10]$): ($1/\beta$ indicated with a dot)



The exponential exponential family (sic) is widely used, for example when modeling the time between independent events [Johnson & Kotz 1970, §18.2$_{208}$]. Each member has the set of all finite time differences $\mathscr{X} := \mathbb{R}_{\geq 0}$ as a sample space and is parameterized by the rate of the occurrence of events, i.e., a $\phi := \beta$ in $\Phi := \mathbb{R}_{>0}$. Let f be a measurable gamble on $\mathbb{R}_{\geq 0}$, the exponential linear prevision is then defined by

$$\mathrm{Ex}(f|\beta) := \int_{\mathbb{R}_{\geq 0}} f z\cdot\beta\cdot\exp(-\beta\cdot z)\,\mathrm{d}z. \tag{B.32}$$

Writing the likelihood in exponential family form is trivial: let z be an observed time difference, then the likelihood can be written as

$$\beta\cdot\exp(-\beta\cdot z) = \beta\cdot\exp\langle -\beta|z\rangle.$$

Comparing this last expression with (4.1)$_{156}$, we see that the Euclidean vector space for this family has dimension $d := 1$ and that

$$\begin{aligned} &\tau := \mathrm{id}_{\mathbb{R}_{\geq 0}}, \quad \text{so} \quad \mathscr{T} := \mathrm{co}\,\mathbb{R}_{\geq 0} = \mathbb{R}_{\geq 0},\\ &\psi := \beta : \mathbb{R}_{>0} ; -\beta, \quad \text{so} \quad \Xi := \mathbb{R}_{<0}, \end{aligned} \tag{B.33}$$

(so this is a linear exponential family) and also that

$$\mathsf{a} := \mathbb{R}_{\geq 0} ; 1 \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R}_{<0} ; -\xi. \tag{B.34}$$

The cumulant function for the exponential exponential family becomes

$$\kappa := -\ln \circ \mathsf{b} = \xi : \mathbb{R}_{<0} ; -\ln(-\xi).$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find – letting $\xi := \psi\beta$ – that

$$\mathrm{Ex}(\tau|\beta) = (\mathrm{D}\kappa)\xi = \tfrac{1}{\beta}, \tag{B.35}$$

which is the mean time between event occurrences.

The information in (B.33) and (B.34) fully defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}_{<0}$ and the domain of possible parameters is $\operatorname{int}\mathscr{T} = \mathbb{R}_{>0}$. The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and t an element of $\operatorname{int}\mathscr{T}$)

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\Xi}(-\xi)^s \cdot \exp^s\langle\xi|t\rangle \,\mathrm{d}\xi.$$

It can be calculated more easily when starting from the original non-canonical parameterization ψ:

$$\mathsf{c}(s,t) = 1/\textstyle\int_{\mathbb{R}_{>0}}\beta^s \cdot \exp(-s\cdot t\cdot\beta)\,\mathrm{d}\beta = \frac{(s\cdot t)^{s+1}}{\Gamma(s+1)}, \tag{B.36}$$

which is easily found after recognizing that the integral is one over the normalization factor of $\mathrm{Ga}(\bullet|s+1, s\cdot t)$ (cf. $(4.31)_{169}$). So the prior parametric linear prevision is just this gamma prevision.

Using $(4.26)_{167}$, (B.36), and (B.34), we can obtain the immediate prior predictive prevision. Its probability density for z in $\mathbb{R}_{\geq 0}$ is

$$\tfrac{(s\cdot t)^{s+1}}{\Gamma(s+1)} \Big/ \tfrac{(s\cdot t+z)^{s+2}}{\Gamma(s+2)} = (s+1)\cdot\tfrac{(s\cdot t)^{s+1}}{(s\cdot t+z)^{s+2}} = \tfrac{s+1}{s\cdot t}\cdot\tfrac{1}{\left(1+\frac{z}{s\cdot t}\right)^{s+2}},$$

for which gamma-compound exponential density would be an apt name. The posterior is obtained by the typical substitution (cf. $§4.1.6_{168}$).

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for exponential sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \mathbb{R}_{>0}$. As this set is already bounded from below, the simplest choice is to only take an upper bound for t, i.e., an upper bound on the mean time between event occurrences.

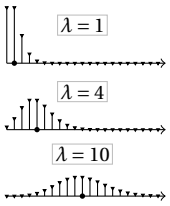
B.2 Discrete families

This section with exponential families on discrete sample spaces contains only one common family, the one formed by the Poisson distributions ($§B.2.1_{\curvearrowright}$). The class of Bernoulli-type discrete families for sampling from finite sets of unordered categories, has already seen ample attention in 'Inference models for exponential families'$_{154}$. The last family

treated here, in §B.2.2, is an artificial one; it is designed out of frustration, to compensate for my all too limited number of encounters with distributions for sampling from finite, structured sets.

B.2.1 *Poisson sampling*

Three Poisson mass functions (plot restricted to 0..20): (the mean is indicated with a dot)



The Poisson family of distributions is widely used, for example when modeling the number of events occurring in a time interval of fixed length [Johnson et al. 2005, §4.9$_{186}$]. Each member has the set of natural numbers $\mathscr{X} := \mathbb{N}$ as a sample space and is parameterized by the mean number of occurrences per interval, i.e., a $\phi := \lambda$ in $\Phi := \mathbb{R}_{>0}$ (0 is excluded for technical reasons). Let f be a gamble on $\mathbb{N}$, the Poisson linear prevision is then defined by

$$\mathrm{Pn}(f|\lambda) := \textstyle\sum_{z:\mathbb{N}} f z \cdot \exp(-\lambda) \cdot \frac{\lambda^z}{z!}. \tag{B.37}$$

Writing the likelihood in exponential family form is straightforward: let z be an observed number of occurrences, then the likelihood can be written as

$$\exp(-\lambda) \cdot \tfrac{\lambda^z}{z!} = \exp(-\lambda) \cdot \exp\langle -\ln\lambda | z\rangle \cdot \tfrac{1}{z!}.$$

Comparing this last expression with (4.1)$_{156}$, we see that the Euclidean vector space for this family has dimension $d := 1$ and that

$$\begin{aligned} \tau &:= \mathrm{id}_{\mathbb{N}}, \quad \text{so} \quad \mathscr{T} := \mathrm{co}\,\mathbb{N} = \mathbb{R}_{\geq 0}, \\ \psi &:= \lambda : \mathbb{R}_{>0} ; \ln\lambda, \quad \text{so} \quad \Xi := \mathbb{R}, \end{aligned} \tag{B.38}$$

(so this is a linear exponential family) and also that

$$\mathsf{a} := z : \mathbb{N} ; \tfrac{1}{z!} \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R} ; \exp(-\exp\xi). \tag{B.39}$$

The cumulant function for the Poisson family becomes

$$\kappa := -\ln \circ \mathsf{b} = \xi : \mathbb{R} ; \exp\xi.$$

So, using (4.7)$_{159}$ and (4.8)$_{159}$, we find – letting $\xi := \psi\lambda$ – that

$$\mathrm{Pn}(\tau|\lambda) = (\mathrm{D}\kappa)\xi = \lambda, \tag{B.40}$$

which is the mean number of occurrences per interval.

The information in (B.38) and (B.39) fully defines the canonical form (4.21)$_{165}$ of the conjugate family prevision: it is defined on all measurable gambles on $\Xi = \mathbb{R}$ and the domain of possible parameters is $\mathrm{int}\,\mathscr{T} = \mathbb{R}_{>0}$. The one thing we do not yet know explicitly is the normalization factor (4.22)$_{165}$: (let s be a strictly positive real number and t an element of $\mathrm{int}\,\mathscr{T}$)

$$\mathsf{c}(s,t) = 1 / \textstyle\int_{\Xi} \exp^s(-\exp\xi) \cdot \exp^s\langle \xi | t\rangle \,\mathrm{d}\xi.$$

As nearly always, it can be calculated more easily when starting from the original noncanonical parameterization ψ:

$$\begin{aligned}\mathrm{c}(s,t) &= 1/\textstyle\int_{\mathbb{R}_{>0}} \exp(-s\cdot\lambda)\cdot\lambda^{s\cdot t}\cdot\left|\left|(\nabla\psi)\lambda\right|\right|\mathrm{d}\lambda \\ &= 1/\textstyle\int_{\mathbb{R}_{>0}} \exp(-s\cdot\lambda)\cdot\lambda^{s\cdot t}\cdot\frac{1}{\lambda}\,\mathrm{d}\lambda \\ &= 1/\textstyle\int_{\mathbb{R}_{>0}} \lambda^{s\cdot t-1}\cdot\exp(-s\cdot\lambda)\,\mathrm{d}\lambda = \frac{s^{s\cdot t}}{\Gamma(s\cdot t)},\end{aligned} \tag{B.41}$$

which is easily found after recognizing that the integral is one over the normalization factor of $\mathrm{Ga}(\bullet|s\cdot t,s)$ (cf. $(4.31)_{169}$). So the prior parametric linear prevision is just this gamma prevision.

Using $(4.26)_{167}$, $(B.36)_{215}$, and $(B.34)_{214}$, we can obtain the immediate prior predictive prevision. Its probability mass for z in $\mathbb{N}$ is

$$\tfrac{1}{z!}\cdot\tfrac{s^{s\cdot t}}{\Gamma(s\cdot t)}\Big/\tfrac{(s+1)^{s\cdot t+z}}{\Gamma(s\cdot t+z)} = \tfrac{\Gamma(s\cdot t+z)}{z!\cdot\Gamma(s\cdot t)}\cdot\tfrac{s^{s\cdot t}}{(s+1)^{s\cdot t+z}} = \tbinom{s\cdot t+z-1}{z}\cdot\left(\tfrac{s}{s+1}\right)^{s\cdot t}\cdot\left(1-\tfrac{s}{s+1}\right)^{z},$$

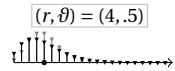
which is the expression for a negative binomial probability mass function. The negative binomial linear prevision [see, e.g., Johnson et al. 2005, §5.1$_{208}$] is defined by (let f be a measurable gamble on $\mathbb{N}$)

$$\mathrm{Nb}(f|r,\vartheta) := \textstyle\sum_{z:\mathbb{N}} f z\cdot\binom{r+z-1}{z}\cdot\vartheta^{r}\cdot(1-\vartheta)^{z}, \tag{B.42}$$

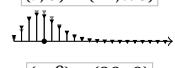
where ϑ in $[0,1]$ is a frequency and r is a strictly positive real number; its mean is $r\cdot\frac{1-\vartheta}{\vartheta}$. (Note that the negative binomial distribution converges to the Poisson distribution for increasing r.) So the immediate prior predictive linear prevision is $\mathrm{Nb}\left(\bullet\;\middle|\; s\cdot t,\frac{s}{s+1}\right)$. The posterior is obtained by the typical substitution (cf. §4.1.6$_{168}$).

Three negative binomial mass functions (plot restricted to 0..20): (the mean is indicated with a dot)

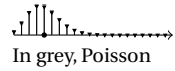
$(r,\vartheta) = (4,.5)$



$(r,\vartheta) = (12,.75)$



$(r,\vartheta) = (36,.9)$



In grey, Poisson mass functions with the same mean are given.

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for Poisson sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\mathrm{int}\,\mathscr{T} = \mathbb{R}_{>0}$. As this set is already bounded from below, the simplest choice is to only take an upper bound for t, i.e., an upper bound on the mean number of occurrences per interval.

B.2.2 *Sampling balanced ternary numbers*

After having browsed through compendia of probability distributions [such as, e.g., Johnson et al. 2005, 1997; Kotz et al. 2000] for a while, it has happened to me multiple times (twice at least) that I wondered: Where are all the probability distributions for finite, structured sets? (Do not read too many technicalities into the word 'structured': I was thinking about things like three collinear or four coplanar points.)

It might be that they are not practically useful, that I have browsed over them, or that they hide in that part of the literature I have not accessed or have no access to. After getting intoxicated by the exponential families, I did not care for these possibilities. I was convinced I could brew my own exponential family and define an inference model for it, which is what we are now going to do.

Let us consider sampling from $\mathscr{X} := \{-1,0,1\}$, the base set for the balanced ternary number system [Wikipedia 2008*a*]. To keep things as simple as possible, our own exponential family – let us call it the Ternouilli family – is going to be the linear canonical one for this sample space. To immediately break our vow of simplicitly – without regrets –, let us take $\mathsf{a} := z : \mathscr{X} \,;\, 1/2^{|z|}$ as the basic mass function (cf. $(\mathrm{i})_{156}$), chosen just because it lets my favorite hyperbolic function appear in the next paragraph. But before leaving this one, it is good to mention that the Ternouilli family likelihood function is proportional to $\exp\langle \bullet | z\rangle \cdot 1/2^{|z|}$ (cf. $(4.1)_{156}$; z is some observed sample from $\{-1,0,1\}$).

As all possible samples are real numbers, this means that the Euclidean vector space for this family has dimension $d := 1$ and so the set of canonical parameters Ξ is a subset of the reals. Recalling first the standard expression for a discrete exponential family prevision $(4.2)_{156}$, for every tentative canonical parameter ξ in $\mathbb{R}$, the normalization factor's value – if defined – is given by (cf. $(\mathrm{v})_{157}$)

$$\mathsf{b}\xi = 1/\textstyle\sum_{z:\{-1,0,1\}} \exp\langle \xi | z\rangle \cdot \mathsf{a}z = \frac{1}{\exp(-\xi)\cdot\frac{1}{2}+1+\exp(-\xi)\cdot\frac{1}{2}} = \frac{1}{1+\cosh\xi}.$$

We see that it is well defined for all ξ in $\mathbb{R}$, so $\Xi := \mathbb{R}$ (cf. $(\mathrm{vi})_{157}$) and the likelihood of ξ is

$$\tfrac{1}{1+\cosh\xi} \cdot \exp\langle \xi | z\rangle \cdot 1/2^{|z|}.$$

Thus for any ξ in Ξ, the Ternouilli linear prevision is defined by (let $f : \mathscr{L}_{\mathbb{R}}$)

$$\mathrm{Tr}(f|\xi) := \textstyle\sum_{z:\{-1,0,1\}} fz \cdot \frac{1}{1+\cosh\xi} \cdot \exp(\xi \cdot z) \cdot \frac{1}{2^{|z|}}. \tag{B.43}$$

To recapitulate, the Ternouilli family's characteristics are

$$\begin{aligned} &\tau := \mathrm{id}_{\{-1,0,1\}}, \quad \text{so} \quad \mathscr{T} := \mathrm{co}\{-1,0,1\} = [-1,1], \\ &\psi := \mathrm{id}_{\mathbb{R}}, \quad \text{so} \quad \Xi := \mathbb{R}, \end{aligned} \tag{B.44}$$

and

$$\mathsf{a} := z : \{-1,0,1\} \,;\, \tfrac{1}{2^{|z|}} \quad \text{and} \quad \mathsf{b} := \xi : \mathbb{R} \,;\, \tfrac{1}{1+\cosh\xi}. \tag{B.45}$$

The cumulant function for the Ternouilli family becomes

$$\kappa := -\ln\circ\mathsf{b} = \xi : \mathbb{R} \,;\, \ln(1+\cosh\xi).$$

So, using $(4.7)_{159}$ and $(4.8)_{159}$, we find that

$$\mathrm{Tr}(\tau|\xi) = (\mathrm{D}\kappa)\xi = \tfrac{\sinh\xi}{1+\cosh\xi}, \tag{B.46}$$

which is just the mean, of course, as the family is linear. This indicates that working with the mean parameterization (i.e., $\psi := \mu :]-1,1[\,;\ln\frac{1+\mu}{1-\mu}$) might have been more intuitive. Ah, well; as we say in Dutch: gedane zaken nemen geen keer.

The information in (B.44) and (B.45) fully defines the canonical form $(4.21)_{165}$ of the conjugate family prevision: it is defined on all measur-

able gambles on $\Xi = \mathbb{R}$ and the domain of parameters is $\operatorname{int}\mathscr{T} =]-1,1[$. The one thing we do not yet know explicitly is the normalization factor $(4.22)_{165}$: (let s be a strictly positive real number and $t : \operatorname{int}\mathscr{T}$)

$$\begin{aligned}\mathrm{c}(s,t) &= 1/\textstyle\int_{\Xi}\big(\frac{1}{1+\cosh\xi}\big)^{s}\cdot\exp^{s}\langle\xi\,|\,t\rangle\,\mathrm{d}\xi\\ &= 1/\textstyle\int_{\mathbb{R}}\big(\frac{1}{1+\cosh\xi}\big)^{s}\cdot\exp^{s}(\xi\cdot t)\,\mathrm{d}\xi;\end{aligned}$$

which, using the substitution $\zeta := \exp\xi$, becomes

$$\begin{aligned}&= 1/\textstyle\int_{\mathbb{R}_{>0}} 2^{s}\cdot\frac{\zeta^{s\cdot t-1}}{(2+\zeta+\zeta^{-1})^{s}}\,\mathrm{d}\zeta\\ &= 2^{-s}/\textstyle\int_{\mathbb{R}_{>0}}\frac{\zeta^{s\cdot(t+1)-1}}{(1+\zeta)^{2\cdot s}}\,\mathrm{d}\zeta = \frac{2^{-s}\cdot\Gamma(2\cdot s)}{\Gamma(s\cdot(1+t))\cdot\Gamma(s\cdot(1-t))},\end{aligned}\tag{B.47}$$

where the last step follows from spotting the beta integral [Abramowitz & Stegun 1972, §6.2$_{258}$].

So the prior parametric linear prevision is (let $f : \mathscr{L}_{\mathbb{R}}$)

$$\mathrm{Tc}(f\,|\,s,t) := \textstyle\int_{\mathbb{R}} f\xi\cdot\frac{2^{-s}\cdot\Gamma(2\cdot s)}{\Gamma(s\cdot(1+t))\cdot\Gamma(s\cdot(1-t))}\cdot\big(\frac{1}{1+\cosh\xi}\big)^{s}\cdot\exp^{s}(\xi\cdot t)\,\mathrm{d}\xi.\tag{B.48}$$

Using $(4.27)_{167}$, (B.47), and (B.45), we can obtain the immediate prior predictive prevision. Its probability mass for z in $\{-1,0,1\}$ is

$$\begin{aligned}&\frac{\frac{1}{2^{|z|}}\cdot\frac{2^{-s}\cdot\Gamma(2\cdot s)}{\Gamma(s\cdot(1+t))\cdot\Gamma(s\cdot(1-t))}}{\frac{2^{-(s+1)}\cdot\Gamma(2\cdot(s+1))}{\Gamma((s+1)\cdot(1+\frac{s\cdot t+z}{s+1}))\cdot\Gamma((s+1)\cdot(1-\frac{s\cdot t+z}{s+1}))}}\\ &\qquad= \frac{2}{2^{|z|}}\cdot\frac{\Gamma(2\cdot s)}{\Gamma(2\cdot s+2)}\cdot\frac{\Gamma(s\cdot(1+t)+1+z)}{\Gamma(s\cdot(1+t))}\cdot\frac{\Gamma(s\cdot(1-t)+1-z)}{\Gamma(s\cdot(1-t))};\end{aligned}$$

we can rewrite this by using generalized binomial coefficients:

$$\begin{aligned}&= \frac{2}{2^{|z|}}\cdot\frac{1}{(2\cdot s+1)\cdot(2\cdot s)}\cdot\tbinom{s\cdot(1+t)+z}{1+z}\cdot(1+z)!\cdot\tbinom{s\cdot(1-t)-z}{1-z}\cdot(1-z)!\\ &= \frac{1}{(2\cdot s+1)\cdot s}\cdot\tbinom{s\cdot(1+t)+z}{1+z}\cdot\tbinom{s\cdot(1-t)-z}{1-z}.\end{aligned}$$

The posterior is obtained by the typical substitution (cf. §4.1.6$_{168}$).

Now all it takes to define imprecise-probabilistic inference models of the type of (4.40)–$(4.43)_{174}$ for Ternouilli sampling is to choose a number of pseudocounts s in $\mathbb{R}_{>0}$ and a bounded subset of $\operatorname{int}\mathscr{T} = \operatorname{int}\mathscr{T} =]-1,1[$. As this is already a bounded set, the easiest choice is to take $\mathscr{U} := \operatorname{int}\mathscr{T}$ and start with a near-ignorance prior, unless there is so much prior information about the Ternouilli process that would make it worthwhile to specify bounds.

This subsection shows that, if none of the classical sampling models are a good fit for some problem, it might be worth a try to see if a custom-made exponential family sampling model and the corresponding (imprecise-probabilistic) inference models can be constructed.

LOWER & UPPER COVARIANCE

> Die allgemeinen Naturgesetze sind durch Gleichungen auszudrücken, die für alle Koordinatensysteme gelten, d. h. die beliebigen Substitutionen gegenüber kovariant (allgemein kovariant) sind.
>
> Einstein [1916, p.776]

Quaeghebeur [2008] communicates this research to the community.

The variance and covariance of random variables are important concepts in classical statistics. Walley [1991, §G_{617}] proposes a generalization to the theory of coherent lower previsions of the classical definition of variance. In this appendix, which stands somewhat apart from the rest of this thesis, we add a generalization of the classical definition of covariance. (None of the results here are used elsewhere.)

G.0.3 *Lower & upper variance*

Let us first sketch Walley's [1991, §$G1_{617}$] approach to defining lower and upper variance.

Given a possibility space Ω and a linear prevision P on $\mathscr{L}_\Omega$, the variance $\mathrm{var}_P : \mathscr{L}_\Omega \to \mathbb{R}$ is the function defined for every gamble f on Ω by

$$\begin{aligned} \mathrm{var}_P f &= P(f - Pf)^2 \\ &= Pf^2 - (Pf)^2. \end{aligned} \tag{G.1}$$

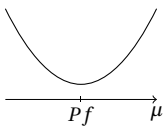
Taking into account the topological assumptions made at the beginning of §1.2.9_{50}, the second form makes it clear that, for any gamble f on Ω, the function $P : (\mathscr{P}\mathscr{L}_\Omega)_{\mathrm{lin}} ; \mathrm{var}_P f$ is continuous.

This definition can be rewritten as follows: for all real μ it holds that

$$\mathrm{var}_P f = P(f - \mu + \mu - Pf)^2$$

and thus – because of the linearity of P – that

$$P(f-\mu)^2 = \mathrm{var}_P f + (Pf - \mu)^2,$$



Note that the second term in the right-hand side as a function of μ is a parabola with minimum zero in Pf; the left-hand side is therefore also an expression for a parabola, with the variance as a minimum, also attained in Pf. This leads to an alternative definition of the variance:

$$\mathrm{var}_P f = \min_{\mu:\mathbb{R}} P(f-\mu)^2. \tag{G.2}$$

This alternative definition gives us the inspiration we need for the definition of lower and upper variance. To wit, for any given coherent

lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$, the lower variance $\underline{\mathrm{var}}_{\underline{P}} : \mathscr{L}_\Omega \to \mathbb{R}$ and upper variance $\overline{\mathrm{var}}_{\underline{P}} : \mathscr{L}_\Omega \to \mathbb{R}$ are defined by

$$\underline{\mathrm{var}}_{\underline{P}} f = \min_{\mu:\mathbb{R}} \underline{P}(f-\mu)^2, \tag{G.3}$$

$$\overline{\mathrm{var}}_{\underline{P}} f = \min_{\mu:\mathbb{R}} \overline{P}(f-\mu)^2. \tag{G.4}$$

These definitions for lower and upper variance and the rewritten definition of variance itself can be seen as optimization problems.

We must still justify using minimum operators and not infimum operators in (G.3) and (G.4). To this end, let $\overline{\underline{P}} : \{\underline{P}, \overline{P}\}$, so we can do a parallel derivation for both cases. Furthermore, let $\varepsilon := \mu - \underline{P} f$, then we can write

$$\begin{aligned}
\overline{\underline{P}}(f-\mu)^2 &= \overline{\underline{P}}(f - \underline{P}f - \varepsilon)^2 \\
&\geq \overline{\underline{P}}(f-\underline{P}f)^2 + \varepsilon^2 + \underline{P}(-2\cdot\varepsilon\cdot(f-\underline{P}f)) \\
&\geq \overline{\underline{P}}(f-\underline{P}f)^2 + \varepsilon^2 \quad \text{if } \varepsilon \leq 0 \\
&> \overline{\underline{P}}(f-\underline{P}f)^2 \quad \text{if } \varepsilon < 0, \text{ i.e., if } \mu < \underline{P}f.
\end{aligned} \tag{G.5}$$

The first inequality follows from superadditivity $(1.31)_{42}$ or mixed superadditivity, the conjugate version of mixed subadditivity $(1.36)_{42}$. An entirely similar derivation, now with $\varepsilon := \overline{P} f - \mu$, gives us

$$\overline{\underline{P}}(f-\mu)^2 > \overline{\underline{P}}(f-\overline{P}f)^2 \quad \text{if } \varepsilon < 0, \text{ i.e., if } \mu > \overline{P}f. \tag{G.6}$$

Together with the fact that $[\underline{P}f, \overline{P}f]$ is compact and that $\overline{\underline{P}}(f-\mu)^2$ is a continuous function of μ, (G.5) and (G.6) show that, for both the case of the lower and upper variance, a minimum is attained in a μ that belongs to $[\underline{P}f, \overline{P}f]$.

With his variance envelope theorem, Walley [1991, §G2_{618}] shows that this definition is equivalent to the one that would be obtained by taking the lower and upper variance as the lowest and highest variance attained by the elements of the credal set $\mathscr{M}\underline{P}$. Explicitly, he proves that

$$\underline{\mathrm{var}}_{\underline{P}} f = \min_{P:\mathscr{M}\underline{P}} \mathrm{var}_P f, \tag{G.7}$$

$$\overline{\mathrm{var}}_{\underline{P}} f = \max_{P:\mathscr{M}\underline{P}} \mathrm{var}_P f. \tag{G.8}$$

Considering the equivalence as uncertainty model of a coherent lower prevision and its credal set, it would be nice if we could let this kind of result also hold for a generalized definition of covariance. This will indeed be the case; but, in the next subsection, we first try to propose a direct definition, i.e., not in terms of the credal set.

G.0.4 *Covariance as an optimization problem*

The approach to obtaining a generalized definition for covariance runs parallel to the one used for variance.

We again start from the classical definition. So, consider a possibility space Ω and a linear prevision P on $\mathscr{L}_\Omega$, the covariance $\mathrm{cov}_P : (\mathscr{L}_\Omega)^2 \to \mathbb{R}$ is the function defined for every pair of gambles f and g on Ω by

$$\begin{aligned}\mathrm{cov}_P(f,g) &= P\big((f-Pf)\cdot(g-Pg)\big)\\ &= P(f\cdot g) - Pf\cdot Pg.\end{aligned} \tag{G.9}$$

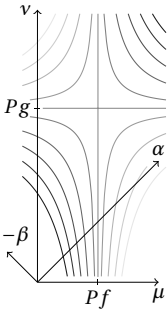
Again, the second form makes it clear that, for any pair of gambles f and g on Ω, the function $P : (\mathscr{P}\mathscr{L}_\Omega)_{\mathrm{lin}} ; \mathrm{cov}_P(f,g)$ is continuous.

We again rewrite this as an optimization problem: for all real μ and ν

$$\mathrm{cov}_P(f,g) = P\big((f-\mu+\mu-Pf)\cdot(g-\nu+\nu-Pg)\big)$$

and thus

$$P\big((f-\mu)\cdot(g-\nu)\big) = \mathrm{cov}_P(f,g) + (Pf-\mu)\cdot(Pg-\nu).$$



(Darker is higher.)

The second term of the right-hand side is the expression in μ,ν for a hyperbolic paraboloid (or saddle surface); the same therefore again holds for the left-hand side. Its saddle point $\underline{P}f, \underline{P}g$ can be reached using a minimax (or maximin) operator. This is clearer after a substitution and some rewriting: let $\alpha,\beta : \mathbb{R}^2$ be such that $\mu = \alpha+\beta$ and $\nu = \alpha-\beta$, then it holds for all $\alpha,\beta : \mathbb{R}^2$ that

$$P\big((\tfrac{f+g}{2}-\alpha)^2 - (\tfrac{f-g}{2}-\beta)^2\big) = \mathrm{cov}_P(f,g) + (P\tfrac{f+g}{2}-\alpha)^2 - (P\tfrac{f-g}{2}-\beta)^2,$$

which gives rise to two *equivalent* definitions:

$$\mathrm{cov}_P(f,g) = \min_{\alpha:\mathbb{R}}\max_{\beta:\mathbb{R}} P\big((\tfrac{f+g}{2}-\alpha)^2 - (\tfrac{f-g}{2}-\beta)^2\big), \tag{G.10}$$

$$\mathrm{cov}_P(f,g) = \max_{\beta:\mathbb{R}}\min_{\alpha:\mathbb{R}} P\big((\tfrac{f+g}{2}-\alpha)^2 - (\tfrac{f-g}{2}-\beta)^2\big). \tag{G.11}$$

Because P is linear and its argument is the sum of terms in either α or β, it is the same whether we use a maximin or minimax operator.

Proposing a definition for lower covariance $\underline{\mathrm{cov}}_{\underline{P}} : (\mathscr{L}_\Omega)^2 \to \mathbb{R}$ and upper covariance $\overline{\mathrm{cov}}_{\underline{P}} : (\mathscr{L}_\Omega)^2 \to \mathbb{R}$ would ideally have consisted of just replacing the linear prevision P on $\mathscr{L}_\Omega$ with some coherent lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$ (or its conjugate upper prevision $\overline{P}$). However, the fact that we have two operators to choose from leaves us with a dilemma: as neither $\underline{P}$ or $\overline{P}$ is linear, does it matter which operator to use for the definition of lower and upper covariance? Perhaps working with credal sets can shed some light on this issue and clarify which of the two choices should be taken (if they do not turn out to be equivalent). This is the topic of the next subsection.

However, because it follows from both (G.10) and (G.11) and the linearity of P that

$$\begin{aligned}\mathrm{cov}_P(f,g) &= \min_{\alpha:\mathbb{R}} P(\tfrac{f+g}{2}-\alpha)^2 - \min_{\beta:\mathbb{R}} P(\tfrac{f-g}{2}-\beta)^2\\ &= \mathrm{var}_P\,\tfrac{f+g}{2} - \mathrm{var}_P\,\tfrac{f-g}{2},\end{aligned} \tag{G.12}$$

there is one thing we can already say: Independently of the operator, using the same reasoning that led to $(G.5)_{221}$ and $(G.6)_{221}$, it follows that the minimizing α belongs to $[\underline{P}\frac{f+g}{2}, \overline{P}\frac{f+g}{2}]$ and – after invoking conjugacy – that the maximizing β belongs to $[\underline{P}\frac{f-g}{2}, \overline{P}\frac{f-g}{2}]$.

G.0.5 *The covariance envelope theorem*

Even though we do not yet have a direct definition of lower and upper covariance, an indirect definition is given by requiring the covariance envelope theorem to hold. To wit, that for any possibility space Ω, any coherent lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$, and any two gambles f and g on Ω it holds that

$$\underline{\mathrm{cov}}_{\underline{P}}(f,g) = \min_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g), \tag{G.13}$$

$$\overline{\mathrm{cov}}_{\underline{P}}(f,g) = \max_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g). \tag{G.14}$$

Due to the compactness of $\mathscr{M}\underline{P}$ and the continuity of the covariance as a function of P, the minimum and maximum are attained, so this theorem is sensible as an indirect definition. The minimum and maximum above are not necessarily attained in an extreme point of $\mathscr{M}\underline{P}$, in contrast to the situation for lower previsions $(1.55)_{50}$.

In the proof of Walley's [1991, §G2$_{618}$] variance envelope theorem, a similar interchange – now with just a minimization operator – takes a central role.

Looking back at the two equivalent definitions (G.10) and (G.11) of covariance, it becomes clear we must investigate whether the maximin and minimax operator encountered there can be interchanged with the maximum or minimum over P encountered in (G.13) and (G.14). Let us write this more explicitly. First define the convex functions

$$u := \alpha : \mathbb{R} ; (\tfrac{f+g}{2} - \alpha)^2, \qquad v := \beta : \mathbb{R} ; (\tfrac{f-g}{2} - \beta)^2,$$

then the question is: Which, if any, of the following statements can we ascertain to be true:

$$\min_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g) = \min_{\alpha:\mathbb{R}} \max_{\beta:\mathbb{R}} \min_{P:\mathscr{M}\underline{P}} P(u\alpha - v\beta),$$

$$\min_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g) = \max_{\beta:\mathbb{R}} \min_{\alpha:\mathbb{R}} \min_{P:\mathscr{M}\underline{P}} P(u\alpha - v\beta),$$

$$\max_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g) = \min_{\alpha:\mathbb{R}} \max_{\beta:\mathbb{R}} \max_{P:\mathscr{M}\underline{P}} P(u\alpha - v\beta),$$

$$\max_{P:\mathscr{M}\underline{P}} \mathrm{cov}_P(f,g) = \max_{\beta:\mathbb{R}} \min_{\alpha:\mathbb{R}} \max_{P:\mathscr{M}\underline{P}} P(u\alpha - v\beta)?$$

First of all, note that consecutive minimum operators or consecutive maximum operators can always be interchanged.

Whether the interchange of a minimum and a maximum operator is allowed, can only be checked after a more thorough study of the functions involved: As function application is a linear operation, $P(u\alpha - v\beta)$ is linear in P (and thus both convex and concave); as P is linear, $P(u\alpha - v\beta)$ is convex in α and concave in β. Furthermore, $P(u\alpha - v\beta)$ is continuous in α, β, and P. Together with the fact that the maximum or minimum is always attained in some convex compact set, this is enough to do a first operator interchange, i.e., we can apply the minimax theorem [Walley

1991, §E6$_{613}$] if needed. We find the following modified statements:

$$\begin{aligned}
\min_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \min_{\alpha:\mathbb{R}} \min_{P:\mathscr{M}\underline{P}} \max_{\beta:\mathbb{R}} P(u\alpha - v\beta),\\
\min_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \max_{\beta:\mathbb{R}} \min_{P:\mathscr{M}\underline{P}} \min_{\alpha:\mathbb{R}} P(u\alpha - v\beta),\\
\max_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \min_{\alpha:\mathbb{R}} \max_{P:\mathscr{M}\underline{P}} \max_{\beta:\mathbb{R}} P(u\alpha - v\beta),\\
\max_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \max_{\beta:\mathbb{R}} \max_{P:\mathscr{M}\underline{P}} \min_{\alpha:\mathbb{R}} P(u\alpha - v\beta).
\end{aligned}$$

As maximizing a convex operation and minimizing is a concave operation, $\max_{\beta:\mathbb{R}} P(u\alpha - v\beta)$ is convex as a function of P and α and $\min_{\alpha:\mathbb{R}} P(u\alpha - v\beta)$ is concave as a function of P and β. This means that a second application of the maximin theorem is not possible, and a second interchange is not generally possible for all cases. We can therefore only be sure about the first and last of the initial statements; to wit,

$$\begin{aligned}
\min_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \min_{\alpha:\mathbb{R}} \max_{\beta:\mathbb{R}} \underline{P}(u\alpha - v\beta),\\
\max_{P:\mathscr{M}\underline{P}} \operatorname{cov}_P(f,g) &= \max_{\beta:\mathbb{R}} \min_{\alpha:\mathbb{R}} \overline{P}(u\alpha - v\beta).
\end{aligned}$$

Thus, the covariance envelope theorem implies a direct definition of lower and upper covariance, the starting point of the next subsection.

G.0.6 *Definition, a property & discussion*

Given some possibility space Ω and a coherent lower prevision $\underline{P}$ on $\mathscr{L}_\Omega$, the lower and upper covariance $\underline{\operatorname{cov}}_{\underline{P}} : (\mathscr{L}_\Omega)^2 \to \mathbb{R}$ and $\overline{\operatorname{cov}}_{\underline{P}} : (\mathscr{L}_\Omega)^2 \to \mathbb{R}$ are defined – for any pair of gambles f and g on Ω – by

$$\underline{\operatorname{cov}}_{\underline{P}}(f,g) = \min_{\alpha:\mathbb{R}} \max_{\beta:\mathbb{R}} \underline{P}\big((\tfrac{f+g}{2} - \alpha)^2 - (\tfrac{f-g}{2} - \beta)^2\big), \tag{G.15}$$

$$\overline{\operatorname{cov}}_{\underline{P}}(f,g) = \max_{\beta:\mathbb{R}} \min_{\alpha:\mathbb{R}} \overline{P}\big((\tfrac{f+g}{2} - \alpha)^2 - (\tfrac{f-g}{2} - \beta)^2\big). \tag{G.16}$$

Note that this definition reduces to the one for lower and upper variance when $g = f$; it also reduces to the classical definition for linear previsions, because they are, well, linear.

An interesting property of classical covariance is that it can be written as a difference of two variances (G.12)$_{222}$. For our generalized definition, this identity becomes a string of inequalities:

$$\begin{aligned}
\underline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2}) &\leq \underline{\operatorname{cov}}_{\underline{P}}(f,g)\\
&\leq \min\big\{\underline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \underline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2}), \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2})\big\}\\
&\leq \max\big\{\underline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \underline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2}), \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2})\big\}\\
&\leq \overline{\operatorname{cov}}_{\underline{P}}(f,g) \leq \overline{\operatorname{var}}_{\underline{P}}(\tfrac{f+g}{2}) - \underline{\operatorname{var}}_{\underline{P}}(\tfrac{f-g}{2}).
\end{aligned} \tag{G.17}$$

These inequalities are obtained starting from the definitions of lower and upper covariance (G.15) and (G.16). They are related to lower and upper variance (G.3)$_{221}$ and (G.4)$_{221}$ by using superadditivity (1.31)$_{42}$ and mixed subadditivity (1.36)$_{42}$.

The derivation given in this section is an interesting example of the use of the properties of coherent lower previsions and of their credal set. However, something is still lacking: What is the *meaning* of lower and upper variance and lower and upper covariance? An intuitive interpretation is the one typically given to their precise counterparts:

(i) variance is a statistic describing a belief about the 'spread' of a gamble,

(ii) covariance is a statistic describing a belief about the 'similarity' of two gambles.

I have found no satisfactory behavioral interpretation; they could be seen as prices, but trying to say for what leads to all too convoluted explanations. Perhaps variance and covariance should just be seen as useful for the description of probability density or mass functions, and any 'generalization' as mathematically interesting at most.

The entries in this bibliography are ordered as follows:

(i) alphabetically (ignoring spaces and capitalization, so 'Cozman' before 'de Campos' before 'Dempster' before 'De Roeck'), then
(ii) by year, and then
(iii) by title.

❧

In many entries, we give a uniform resource locator (URL). All these URLs begin with the same scheme name and delimiting characters, which we drop; so 'URL: www.UGent.be' refers to http://www.UGent.be. There are some specific, persistent URL types that get their own identifier:

HDL: The institutional archive of Ghent University stores electronic versions of publications by its students and researchers. It assigns 'handles' to these archived publications, which can be accessed via the same authority, which we also drop; so 'URL: hdl.handle.net/1854/8954' becomes 'HDL: 1854/8954'.

DOI: Most large publishers assign a digital object identifier – a specific handle type starting with '10.' – to their publications. We again drop the authority they have in common; so 'URL: dx.doi.org/10.1002/ajim.4700110207' becomes 'DOI: 10.1002/ajim.4700110207'.

Although DOIs currently seem to be future of electronic referencing, one custom system still remains important:

JSTOR: Quite a number of scholarly journals are archived by the JSTOR organization. Every archived paper is assigned a stable number that allows access via a specific URL. We drop the authority and the database prefix; so 'URL: www.jstor.org/stable/228734' becomes 'JSTOR: 228734'.

❧

In the margin, we give the page(s) where the reference is cited.

135, 210, 213, 219 Abramowitz M. & Stegun I.A., eds. [1972]. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical tables*. URL: www.math.sfu.ca/~cbm/aands/.

165 Arnold B.C., Castillo E. & Sarabia J.M. [1993]. Conjugate exponential family priors for exponential family likelihoods. *Statistics*, 25: 71–77.

186 Augustin T. & Coolen F.P.A. [2004]. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2): 251–272. DOI: 10.1016/j.jspi.2003.07.003.

70, 86 Avis D. [2000]. lrs: A Revised Implementation of the Reverse Search Vertex Enumeration Algorithm. In *Polytopes - Combinatorics and Computation*. Ed. by Kalai & Ziegler, 177–198. DMV Seminar 29. Birkhauser-Verlag. URL: cgm.cs.mcgill.ca/~avis/C/lrs.html.

Avis D., Bremner D. & Seidel R. [1997]. How good are convex hull algorithms? *Computational Geometry*, 7: 265–301. DOI: 10.1016/s0925-7721(96)00023-5. 70

Barndorff-Nielsen O. [1978]. *Information and Exponential Families in Statistical Theory*. Wiley. 155–157, 159, 212

Berger J. [1993]. *An Overview of Robust Bayesian Analysis*. Tech. rep. 93-53C. Purdue University, Department of Statistics. URL: www.stat.duke.edu/~berger/papers/overview.html. 227

Berger J. et al. [1994]. An Overview of Robust Bayesian Analysis. *TEST*, 3(1): 5–124. With discussion. DOI: 10.1007/bf02562676. Available without discussion as [Berger 1993]. 141

Bernard J.-M. [1997]. Bayesian analysis of tree-structured data. *Revue Internationale de Systémique*, 11(1): 11–29. 138

Bernard J.-M., Seidenfeld T. & Zaffalon M., eds. [2003]. *ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*. Lugano. Proceedings in Informatics 18. Waterloo, Ontario, Canada: Carleton Scientific. URL: www.sipta.org/isipta03. 232, 234

Bernardo J. M. & Smith A. F. M. [1994]. *Bayesian theory*. Wiley. This book is meant to take up the story where de Finetti [1974–1975] left off. 115–117, 140, 154, 159, 169, 170, 209

Bernstein D. S. [2005]. *Matrix Mathematics*. Princeton University Press. 171, 207

Bierce A. [1911]. *The Devil's Dictionary*. URL: www.gutenberg.org/etext/972. 28

Billingsley P. [1961]. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, 32(1): 12–40. JSTOR: 2237603. 149

Boratyńska A. [1997]. Stability of Bayesian inference in exponential families. *Statistics & Probability Letters*, 36(2): 173–178. DOI: 10.1016/s0167-7152(97)00060-6. 175

Boute R. T. [2005]. Functional Declarative Language Design and Predicate Calculus: A Practical Approach. *ACM Transactions on Programming Languages and Systems*, 27(5): 988–1047. DOI: 10.1145/1086642.1086647. 23

Boyd S. & Vandenberghe L. [2004]. *Convex Optimization*. Cambridge University Press. URL: www.stanford.edu/~boyd/cvxbook/. 178

Briggs P. [2005]. *Ethiopia*. 4th ed. Bradt Travel Guides Ltd. 194

142 Brown G.W. [1951]. Iterative Solutions of Games by Fictitious Play. In *Activity Analysis of Production and Allocation*. Ed. by Koopmans, 374–376. Cowles commission monographs 13. Wiley.

155 Brown L.D. [1986]. *Fundamentals of Statistical Exponential Families*. Hayward, California: Institute of Mathematical Statistics.

72 Brüning M. & Dennenberg D. [2008]. The extreme points of the set of belief measures. *International Journal of Approximate Reasoning*, 48(3): 670–675. DOI: 10.1016/j.ijar.2006.11.003. Explicitation of an implicit result of Choquet [1954].

54, 63, 156, 165, 168 Burrill C.W. [1972]. *Measure, Integration, and Probability*. McGraw-Hill.

118 Carnap R. [1952]. *The Continuum of Inductive Methods*. University of Chicago Press.

73, 87, 93 Chateauneuf A. & Jaffray J.-Y. [1989]. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, 17: 263–283. DOI: 10.1016/0165-4896(89)90056-5.

228 Choquet G. [1954]. Theory of Capacities. *Annales de l'Institut Fourier*, 5: 131–295. URL: www.numdam.org/item?id=aif_1954__5__131_0.

101 Cifarelli D.M. & Regazzini E. [1996]. De Finetti's Contribution to probability and Statistics. *Statistical Science*, 11(4): 253–282. DOI: 10.1214/ss/1032280303.

175 Coolen F.P.A. [1993]. Imprecise conjugate prior densities for the one-parameter exponential family of distributions. *Statistics & Probability Letters*, 16(5): 337–342. DOI: 10.1016/0167-7152(93)90066-r.

186 Coolen F.P.A. & Augustin T. [2009]. A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories. *International Journal of Approximate Reasoning*. DOI: 10.1016/j.ijar.2008.03.011. In press.

229, 234 Cozman F.G., Nau R. & Seidenfeld T., eds. [2005]. *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*. Pittsburgh. URL: www.sipta.org/isipta05.

159, 165 Cramér H. [1946]. *Mathematical methods of statistics*. Princeton University Press.

146 Dayhoff M.O., Schwartz R.M. & Orcutt B.C. [1978]. A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*. Ed. by Dayhoff. Chap. 22, 345–352. National Biomedical Research Foundation.

de Campos L. M., Huete J. F. & Moral S. [1994]. Probability Intervals: A Tool for Uncertain Reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2): 167–196. DOI: 10.1142/s0218488594000146. 73

De Cooman G., Hermans F. & Quaeghebeur E. [2008*a*]. Sensitivity analysis for finite Markov chains in discrete time. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. Helsinki. Ed. by McAllester & Myllymäki, 129–136. HDL: 1854/11638. 146, 152, 229

— [2009*a*]. Imprecise Markov chains and their limit behavior. URL: arxiv.org/abs/0801.0980. Submitted. This is an expanded and improved version of [De Cooman, Hermans & Quaeghebeur 2008*a*]. 147, 152

De Cooman G. & Miranda E. [2007]. Symmetry of models versus models of symmetry. In *Probability and Inference: Essays in Honor of Henry E. Keyburg, Jr.* Ed. by Harper & Wheeler, 67–149. London: King's College Publications. 69, 83, 190

De Cooman G., Miranda E. & Quaeghebeur E. [2007*a*]. Immediate prediction under exchangeability and represenation insensitivity. In [De Cooman, Zaffalon & Vejnarová 2007*c*], 107–116. HDL: 1854/7630. 118

— [2009*b*]. Representation insensitivity in immediate prediction under exchangeability. *International Journal of Approximate Reasoning*. DOI: 10.1016/j.ijar.2008.03.010. In press. 118

De Cooman G., Quaeghebeur E. & Miranda E. [2007*b*]. Representing and assessing exchangeable lower previsions. In *Bulletin of the International Statistical Institute 56th Session – Proceedings*. Lisboa. 1556. HDL: 1854/8320. 95

— [2009*c*]. Exchangeable lower previsions. URL: arxiv.org/abs/0801.1265. Submitted. 95, 103

De Cooman G., Troffaes M. C. M. & Miranda E. [2005*b*]. n-Monotone lower previsions and lower integrals. In [Cozman, Nau & Seidenfeld 2005], 145–154. Expanded versions: [De Cooman, Troffaes & Miranda 2005*a*, 2008*b*]. 73

— [2005*a*]. n-Monotone lower previsions. *Journal of Intelligent & Fuzzy Systems*, 16(4): 253–263. 229

— [2008*b*]. n-Monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347(1): 143–156. DOI: 10.1016/j.jmaa.2008.05.071. URL: arxiv.org/abs/0801.1962. 229

De Cooman G., Zaffalon M. & Vejnarová J., eds. [2007*c*]. *ISIPTA '07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*. Prague. URL: www.sipta.org/isipta07. 229, 235, 236

22,101 de Finetti B. [1937]. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1): 1–68. URL: www.numdam.org/item?id=aihp_1937__7_1_1_0.

230 — [1967]. Quelques conventions qui semblent utiles. *Revue Roumaine des Mathémathiques Pures et Appliquées*, 12: 1227–1233.

230 — [1970]. *Teoria Delle Probabilità*. Giulio Einaudi.

24,31 — [1972*a*]. A Useful Notation. In [de Finetti 1972*b*], xviii–xxiv. Original: [de Finetti 1967].

32,230 — [1972*b*]. *Probability, Induction and Statistics. The art of guessing*. Wiley.

49,107,227 — [1974–1975]. *Theory of Probability*. Trans. from the Italian by A. Machi & A. F. M. Smith. 2 Volumes. Wiley. Original: [de Finetti 1970].

155 DeGroot M. H. [2004]. *Optimal Statistical Decisions*. Wiley Classics Library Edition. Wiley.

72 Dempster A. P. [1967]. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathemathical Statistics*, 38(2): 325–339. DOI: 10.1214/aoms/1177698950. JSTOR: 2239146.

18 De Roeck W., Cox T. & Quaeghebeur E. [2006]. *Verspreking van T. Cox, door W. De Roeck (aangemoedigd door E. Quaeghebeur) omgevormd tot een neologisme*. De achtertuin van het huis in de Zonnebloemstraat te Berchem. June 2006.

147 Dhaenens S. [2007]. Onderzoek van Imprecieze Markov-modellen. MA thesis. Universiteit Gent. HDL: 1854/10573.

212 Dhillon I. S. & Sra S. [2003]. *Modeling Data using Directional Distributions*. Tech. rep. TR-03-06. The University of Texas at Austin. URL: www.cs.utexas.edu/~suvrit/work/research.html.

141,165,168 Diaconis P. & Ylvisaker D. [1979]. Conjugate Priors for Exponential Families. *The Annals of Statistics*, 7(2): 269–281. DOI: 10.1214/aos/1176344611. JSTOR: 2958808.

183 Domingos P. & Pazzani M. [1997]. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29(2): 103–130. DOI: 10.1023/a:1007413511361.

84 Dubois D. & Prade H. [1988]. *Théorie des possibilités*. 2nd ed. Paris: Masson.

154,230 Eco U. [1988]. *Il pendolo di Foucault*. Nederlandse vertaling: [Eco 1989].

230 — [1989]. *De slinger van Foucault*. Nederlandse vertaling van [Eco 1988].

Einstein A. [1916]. Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 49: 769–822. DOI: 10.1002/andp.200590044. URL: www.alberteinstein.info/gallery/gtext3.html. 220

Fine T. L. [1973]. *Theories of Probability*. New York & London: Academic Press. A peculiar rotated i appears on page 4. 32

Fink D. [1995]. *A Compendium of Conjugate Priors*. Tech. rep. Bozeman, MT 59717: Environmental Statistics Group, Department of Biology, Montana State Univeristy. URL: www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf. 154

Fisher N. I., Lewis T. & Embleton B. J. J. [1987]. *Statistical analysis of spherical data*. Cambridge University Press. 210

Friedman J. H. [1997]. On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1(1): 55–77. DOI: 10.1023/A:1009778005914. 183

Friedman J. W. [1989]. *Game Theory with Applications to Economics*. Oxford University Press. 145

Fukuda K. [2004]. *Frequently Asked Questions in Polyhedral Computation*. URL: www.ifor.math.ethz.ch/~fukuda/polyfaq/ (revision of 2004-06-18). 69

Fukuda K. & Prudon A. [1996]. Double Description Method Revisited. In *Combinatorics and Computer Science*. Ed. by Deza, Euler & Manoussakis, 91–111. Lecture Notes in Computer Science 1120. Springer. URL: www.ifor.math.ethz.ch/~fukuda/cdd_home/. 70, 86, 89

Geisser S. [1993]. *Predictive Inference: An Introduction*. Chapman & Hall. 118

Graves R. [1955]. *The Greek Myths*. The quotation refers to the purported ancient Greek technique of sink down bovine sampling of favorable settlement locations. 202

Gutiérrez-Peña E. & Smith A. F. M. [1995]. Conjugate Parameterizations for Natural Exponential Families. *Journal of the American Statistical Association*, 90(432): 1347–1356. JSTOR: 2291525. 165

— [1997]. Exponential and Bayesian Conjugate Families: Review and Extensions. *Test*, 6(1): 1–90. With discussion. 165

Halpern J. Y. & Koller D. [2004]. Representation Dependence in Probabilistic Inference. *Journal of Artificial Intelligence Research*, 21: 319–356. URL: www.jair.org/papers/paper1292.html. 122

Howson C. [2000]. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press. 94

94 Hume D. [1739]. *A treatise of human nature*. URL: www.gutenberg.org/etext/4705.

54 Jaynes E. T. [2003]. *Probability Theory. The Logic of Science*. Ed. by Betthorst. Cambridge University Press.

129 Jeffreys H. [1983]. *Theory of Probability*. 3rd ed. Oxford University Press.

168 Johnson N. L. [1967]. Note on a Uniqueness Relation in Certain Accident Proneness Models. *Journal of the American Statistical Association*, 62(317): 288–289. JSTOR: 2282931.

216, 217 Johnson N. L., Kemp A. W. & Kotz S. [2005]. *Univariate Discrete Distributions*. 3rd ed. Wiley.

213, 214 Johnson N. L. & Kotz S. [1970]. *Continuous Univariate Distributions*. Vol. 1. Boston: Houghton Mufflin Company.

101, 106, 130, 135, 139, 163, 217 Johnson N. L., Kotz S. & Balakrishnan N. [1997]. *Discrete Multivariate Distributions*. Wiley.

237 Johnson W. E. [1924]. *Logic. Part III*. Cambridge University Press.

155 Kallenberg O. [2005]. *Probabilistic Symmetries and Invariance Principles*. Springer.

146 Kemeny J. G. & Snell J. l. [1976]. *Finite Markov Chains*. 2nd ed. Springer.

139, 158, 165, 206, 217 Kotz S., Balakrishnan N. & Johnson N. L. [2000]. *Continuous Multivariate Distributions*. Vol. 1: *Models and Applications*. Wiley.

118 Laplace P.-S. [1825]. *Essai philosophique sur les probabilités*. 5th ed. Paris: Bachelier. URL: books.google.com/books?id=ovo3aaaamaaj.

158 Letac G. [1992]. *Lectures on natural exponential families and their variance functions*. Monografias de Matemática 50. Rio de Janeiro: Conselho Nacional de Desenvolvimento Científico e Tecnológico, Instituto de Matemática Pura e Aplicada (IMPA).

50 Levi I. [1974]. On Indeterminate Probabilities. *The Journal of Philosophy*, 71(13): 391–418.

144 — [1980]. *The Enterprise of Knowledge*. MIT Press.

69, 90 Maaß S. [2003*a*]. Continuous Linear Representations of Coherent Lower Previsions. In [Bernard, Seidenfeld & Zaffalon 2003], 372–382.

50, 69 — [2003*b*]. Exact functionals, functionals preserving linear inequalities, Lévy's metric. PhD thesis. Universität Bremen.

87, 90 — [2005]. *Re: Non-vacuous prevision in extreme points of CLP*($2^{\{1,2,3\}}$). Apr. 6, 2005. Personal communication.

Mardia K.V. & El-Atoum S.A.M. [1976]. Bayesian Inferences for the Von Mises-Fisher Distribution. *Biometrika*, 63(1): 203–206. DOI: 10.1093/biomet/63.1.203. 212

Matheiss T.H. & Rubin D.S. [1980]. A survey and comparison of methods for finding all vertices of convex polyhedral sets. *Mathematics of Operations Research*, 5(2): 167–185. 69

Michie D., Spiegelhalter D.J. & Taylor C.C., eds. [1994]. *Machine Learning, Neural and Statistical Classification*. URL: www.amsta.leeds.ac.uk/~charles/statlog/. 181

Miller R.B. [1980]. Bayesian Analysis of the Two-Parameter Gamma Distribution. *Technometrics*, 22(1): 65–69. JSTOR: 1268384. 214

Miranda E. [2008*b*]. *Updating coherent previsions on finite spaces*. Tech. rep. Department of Statistics & Operations Research, Rey Juan Carlos University. 58

— [2008*a*]. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2): 628–658. DOI: 10.1016/j.ijar.2007.12.001. 49

Miranda E. & De Cooman G. [2007]. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 46(1): 188–225. DOI: 10.1016/j.ijar.2006.12.009. 62, 121

Miranda E., De Cooman G. & Quaeghebeur E. [2006]. The moment problem for finitely additive probabilities. In *Proceedings of the Eleventh International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*. Paris. Vol. 1, 89–96. HDL: 1854/5684. 43

— [2007]. The Hausdorff Moment Problem under Finite Additivity. *Journal of Theoretical Probability*, 20(3): 663–693. DOI: 10.1007/s10959-007-0055-4. HDL: 1854/5684. 43

— [2008*b*]. The moment problem for finitely additive probabilities. In *Uncertainty and Intelligent Information Systems*. Ed. by Bouchon-Meunier, Marsala, Rifqi & Yager. Chap. 3, 33–45. World Scientific. HDL: 1854/12198. URL: www.worldscibooks.com/compsci/6747.html. 43

— [2008*a*]. Finitely additive extensions of distribution functions and moment sequences: the coherent lower prevision approach. *International Journal of Approximate Reasoning*, 48(1): 132–155. DOI: 10.1016/j.ijar.2007.07.007. HDL: 1854/10682. 43, 69

Peterson P.P. [1972]. The Geometry of Radon's Theorem. *The American Mathematical Monthly*, 79(9): 949–963. JSTOR: 2318065. 92

107 Pinkus A. [2005]. Density in Approximation Theory. *Surveys in Approximation Theory*, 1:1–45. URL: www.math.technion.ac.il/sat/papers/1/.

105,106 Prautzsch H., Boehm W. & Paluszny M. [2002]. *Bézier and B-spline Techniques*. Springer.

142 Quaeghebeur E. [2001]. Speltheoretisch leren met imprecieze waarschijnlijkheden: dynamische aspecten. MA thesis. Universiteit Gent. HDL: 1854/6279.

142 — [2003]. Fictitious play: two viewpoints and two versions. In *Proceedings of the 7th workshop on dynamics and computation: Iterated games and cooperation*. Leuven. HDL: 1854/2277.

152 — [2004]. *Weak law for Markov model estimation using the IDM*. Tech. rep. SYSTeMS Research Group, Ghent University. Unpublished, available upon request.

220 — [2008]. Lower & upper covariance. In *Soft Methods for Handling Variability and Imprecision. Proceedings of the 2008 International Conference on Soft Methods in Probability and Statistics*. Toulouse. Ed. by Dubois et al., 323–330. Advances in Soft Computing 48. Springer. DOI: 10.1007/978-3-540-85027-4_39. HDL: 1854/13496.

149 Quaeghebeur E. & De Cooman G. [2003*a*]. Command line completion: an illustration of learning and decision making using the imprecise Dirichlet model. In *Proceedings of the Fourth UGent-FTW PhD Symposium*. Gent. HDL: 1854/2517.

142 — [2003*b*]. Game-Theoretic Learning Using the Imprecise Dirichlet Model. In [Bernard, Seidenfeld & Zaffalon 2003], 452–466. HDL: 1854/2309.

154,186 — [2005]. Imprecise probability models for inference in exponential families. In [Cozman, Nau & Seidenfeld 2005], 287–296. HDL: 1854/3353.

66 — [2006]. Extreme lower probabilities. In *Soft Methods for Integrated Uncertainty Modelling. Proceedings of the 2006 International Workshop on Soft Methods in Probability and Statistics*. Bristol. Ed. by Lawry et al., 211–221. Advances in Soft Computing 6. Springer. DOI: 10.1007/3-540-34777-1_26. HDL: 1854/6276.

66 — [2008]. Extreme lower probabilities. *Fuzzy Sets and Systems*, 159(16): 2163–2175. DOI: 10.1016/j.fss.2007.11.020. HDL: 1854/11713.

142 — [2009]. Learning in games using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*. DOI: 10.1016/j.ijar.2008.03.012. In press.

Quaeghebeur E., De Cooman G. & Aeyels D. [2005]. Building classifiers that cope with small training sets. In *Proceedings of the Sixth UGent-FirW PhD Symposium*. Gent. HDL: 1854/3632. 186

Radiohead [1997]. Karma police. *OK Computer*. URL: www.youtube.com/watch?v=5LeLAELIxKY. 190

Raiffa H. & Schlaifer R. [1968]. *Applied Statistical Decision Theory*. First MIT Press paperback edition. MIT Press. 154

Robinson J. [1951]. An Iterative Method of Solving a Game. *The Annals of Mathematics*, 54(2): 296–301. JSTOR: 1969530. 142

Sarukkai R.R. [2000]. Link prediction and path analysis using Markov chains. *Computer Networks*, 33(1-6): 377–386. DOI: 10.1016/s1389-1286(00)00044-x. 146

Schechter E. [1997]. *Handbook of Analysis and Its Foundations*. Academic Press. 157

Seidenfeld T. & Wasserman L. [1993]. Dilation for sets of probabilities. *The Annals of Statistics*, 21(3): 1139–1154. DOI: 10.1214/aos/1176349254. JSTOR: 2242191. 176

Shafer G. [1976]. *A mathematical theory of evidence*. Princeton University Press. 72, 73

— [1996]. *The Art of Causal Conjecture*. MIT Press. 119

Shannon C.E. [1948]. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27: 379–423, 623–656. URL: cm.bell-labs.com/cm/ms/what/shannonday/paper.html. 46

Shapley L.S. [1971]. Cores of Convex Games. *International Journal of Game Theory*, 1: 11–26. DOI: 10.1007/bf01753431. 92

Škulj D. [2007]. Regular finite Markov chains with interval probabilities. In [De Cooman, Zaffalon & Vejnarová 2007*c*], 405–414. URL: www.sipta.org/isipta07. 147

Smith C.A.B. [1961]. Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society*, B, 23(1): 1–37. JSTOR: 2237603. 33

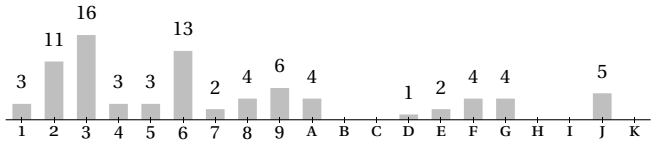
Stendhal [1830]. *Le rouge et le noir*. URL: www.gutenberg.org/etext/798. 18, 66

Troffaes M.C.M. [2005]. Optimality, Uncertainty, and Dynamic Programming with Lower Previsions. PhD thesis. Universiteit Gent. HDL: 1854/5851. 144

— [2007]. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1): 17–29. DOI: 10.1016/j.ijar.2006.06.001. 144, 181

106 Trump W. & Prautzsch H. [1996]. Arbitrarily high degree elevation of Bézier representations. *Computer Aided Geometric Design*, 13: 387–398. DOI: 10.1016/0167-8396(95)00031-3.

22, 23, 28, 31, 32, 34, 36, 38–42, 44, 47–52, 54, 56, 58–64, 70, 76, 78, 80, 94–96, 107, 116, 117, 136, 141, 144, 145, 150, 164, 174–177, 188, 190, 198, 220, 221, 223 Walley P. [1991]. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall. The distribution of the citations over the book's chapters:



95, 123, 136, 138, 140, 154, 175, 192 — [1996]. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society*, B, 58(1): 3–57. With discussion. JSTOR: 2346164.

22, 95, 119, 136, 193 Walley P. & Bernard J.-M. [1999]. *Imprecise Probabilistic Prediction for Categorical Data*. Tech. rep. CAF-9901. Université de Paris 8.

176 Walter G. & Augustin T. [2009]. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*. Conditionally accepted.

180 Walter G., Augustin T. & Peters A. [2007]. Linear Regression Analysis under Sets of Conjugate Priors. In [De Cooman, Zaffalon & Vejnarová 2007*c*], 445–454.

84 Weichselberger K. [2001]. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung*. Vol. I: *Intervallwarscheinlichkeit als umfassendes Konzept*. Unter Mitarbeit von T. Augustin und A. Wallner. Heidelberg: Physica-Verlag.

148 Whittle P. [1955]. Some Distribution and Moment Formulae for the Markov Chain. *Journal of the Royal Statistical Society*, B, 17(2): 235–242. JSTOR: 2983957.

218 Wikipedia [2008*a*]. *Balanced ternary — Wikipedia, The Free Encyclopedia*. URL: en.wikipedia.org/w/index.php?oldid=227595300 (revision of 2008-07-24).

207 — [2008*b*]. *Covariance matrix — Wikipedia, The Free Encyclopedia*. URL: en.wikipedia.org/w/index.php?oldid=232313052 (revision of 2008-08-16).

34 Williams P.M. [1974]. Indeterminate probabilities. In *Formal methods in the methodology of empirical sciences*. Warsaw. Ed. by Przełęcki, Szaniawski & Wójcicki, 229–246. Dordrecht & Boston; Wrocław: D. Reidel Publishing Company; Ossolineum Publishing Company.

52 — [1975]. *Notes on conditional previsions*. Tech. rep. University of Sussex. Published as [Williams 2007].

— [2007]. Notes on Conditional Previsions. *International Journal of Approximate Reasoning*, 44: 366–383. DOI: 10.1016/j.ijar.2006.07.019. 236

Zabell S.L. [1982]. W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, 10(4): 1091–1099. Refers to Johnson [1924]; reprinted in [Zabell 2005]. 125

— [2005]. *Symmetry and Its Discontents: Essay on the History of Inductive Probability*. Cambridge University Press. 237

Zaffalon M. [1999]. A Credal Approach to Naive Classification. In *ISIPTA '99: Proceedings of the First International Symposium on Imprecise probabilities and Their Applications*. Ghent, Belgium. Ed. by De Cooman, Cozman, Moral & Walley, 405–414. URL: www.sipta.org/isipta99. 182

— [2001]. Statistical inference of the naive credal classifier. In *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*. Ed. by De Cooman, Fine & Seidenfeld, 384–393. Ithaca, New York, United States: Shaker Publishing, Maastricht, The Netherlands. URL: www.sipta.org/isipta01. 180, 182, 185, 186

— [2002]. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105: 5–21. DOI: 10.1016/s0378-3758(01)00201-4. 23, 182, 184

— [2005]. Credible classification for environmental problems. *Environmental Modelling & Software*, 20(8): 1003–1012. DOI: 10.1016/j.envsoft.2004.10.006. 182

Zaffalon M., Wesnes K. & Petrini O. [2003]. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine*, 29(1-2): 61–79. DOI: 10.1016/s0933-3657(03)00046-0. 182

Ziegler G.M. [1995]. *Lectures on Polytopes*. Springer. 69

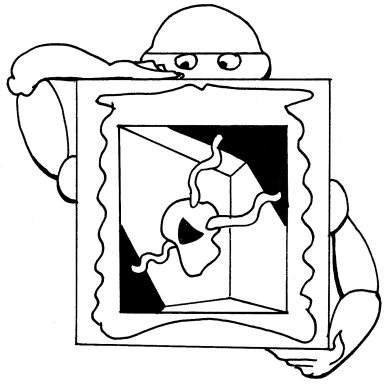
This index of concepts and topics is ordered alphabetically. Page references point to a definition or a first or important use.

This index of symbols is ordered topically. All but the last topic contain symbols for fixed concepts; e.g., the set of reals $\mathbb{R}$ and the generalized gamma function $\Gamma_\bullet$. The last topic – naturally at the back of the index – contains generically used symbols; e.g., Ω for some possibility space and $f_\bullet$ for some function. For functions, the possible arguments and parameters are indicated using the placeholder $\bullet$. Page references, if present, point to a definition or a first or important use.

We have not indexed variants of generic notation created by decorating indexed symbols (using, e.g., primes or bars) or generic symbols that are only used very locally.

Colophon

This thesis was typeset with LaTeX in 10-point Utopia. The memoir class was used, with support from the following main packages: babel, amsmath, microtype, hyperref, and BibLaTeX. The figures were made with Ti*k*Z and the "intervall" program developed by the LMU München. The bibliography and the index were respectively built with the help of BibTeX and makeindex. Last, but not least, invaluable help was found on comp.text.tex.



The artwork in this thesis was graciously provided by Kolja Aertgeerts, who created the black felt-tip pen drawings, & Marjorie Hoefmans, who created the collage and linocut, and also the derivative of each.