On imprecision in statistical theory

by

Marco Y. S. Shum

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Science
in
Statistics

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Name:	Tahani Coolen-Maturi
	Title:	Professor
Supervisors:	Name:	Paul Marriott
	Title:	Professor
	Name:	Tony Wirjanto
	Title:	Professor
Internal members:	Name:	Shoja'eddin Chenouri
	Title:	Professor
	Name:	Martin Lysy
	Title:	Professor
Internal-external Member:	Name:	Thomas Parker
	Title:	Professor

Author's declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis provides an exploration of the interplay between *imprecise probability* and statistics. Mathematically, one may summarise this relationship as how (Bayesian) sensitivity analysis involving a set of (prior) models can be done in relation to the notion of *coherence* in the sense of de Finetti [32], Williams [84] and, more recently, Walley [81]. This thesis explores how imprecise probability can be applied to foundational statistical problems.

The contributions of this thesis are three folds. In Chapter 1, we illustrate and motivate the need for *imprecise models* due to certain inherent limitations of elicitation of a statistical model. In Chapter 2, we provide a primer of imprecise probability aimed at the statistics audience along with illustrative statistical examples and results that highlight salient behaviours of imprecise models from the the statistical perspective.

In the second part of the thesis (Chapters 3, 4, 5), we consider the statistical application of the *imprecise Dirichlet model (IDM)*, an established model in imprecise probability. In particular, the posterior inference for log-odds statistics under sparse contingency tables, the development and use of imprecise interval estimates via quantile intervals over a set of distributions and the geometry of the optimisation problem over a set of distributions are studied. Some of these applications require extensions of Walley's existing framework, and are presented as part of our contribution.

The third part of the thesis (Chapters 6, 7) departs from the IDM parametric assumption and instead focuses on posterior inference using imprecise models in a finite dimensional setting when the lower bound of the probability of the data over a set of elicited priors is zero. This setting generalises the problem of zero marginal probability in Bayesian analysis. In Chapter 6, we explore the methodology, behaviour and interpretability of the posterior inference under two established models in imprecise probability: the *vacuous and regular extensions*. In Chapter 7, we note that these extensions are in fact extremes in *imprecision*, the variability of an inference over the elicited set of probability distributions. Then we consider extensions which are of *intermediate* levels of imprecision, and discuss their elicitation and assessment.

Acknowledgements

Foremost, I am most indebted to my supervisors, Paul Marriott and Tony Wirjanto, for taking this journey with me from the very beginning. Their wisdom has been invaluable for me being able to put so many of the ideas of this thesis into perspective and making them communicable to the reader. They have been instrumental in making concrete my ideas that usually start as a 'stream of consciousness', as Paul would put it. (They have been immensely patient with me regarding that habit.) Above all, their insights about statistics and research have been very influential to me: with due modesty, I would like to think that a modicum of it rubbed off on me, and made me a better statistician. Without doubt, our discussions and interactions have been a joy. Thank you both, sincerely.

I would also like to thank my defence committee for having taken the time to read my thesis and raising inspiring questions. The engaging discussions with them was a pleasure for me.

I would like to thank the following people for their kind support, advice and friendship throughout these years: Joslin Goh, Jay Gweon, David Haskell, Celia Huang, Mirabelle Huynh, Louise Kwan, Garcia (Jiaxi) Liang, Reza Ramezan, Reza Raoufi, Greg Rice, Vincent Russo, Basil Singer, Lu Xin. (I profusely apologise if I have missed anyone.)

I would like to thank the staff members of the Department of Statistics and Actuarial Sciences at UW for their friendly administration of the department. I would like to especially thank Mary Lou Dufton, without whose help I would have otherwise stumbled into all sorts of administrative limbos.

Dedication

To my parents who, under vacuity, had faith in me during this journey.

Table of Contents

List of Tables

List of Figures

Notation

a, $(a_1, \ldots, a_m)$ – a vector in finite dimensional Euclidean space, $\mathbb{R}^m$.

a_C – for a vector $\boldsymbol{a} \in \mathbb{R}^m$, set of indices $C \subseteq \{1, \ldots, m\}$, $a_C = \sum_{i \in C} a_i$.

A^c – the complement of a set A

$\overline{A}$ – the (topological) closure of a set A

$\triangle^m$, $\overline{\triangle^m}$ – the unit simplex (and its topological closure) in $\mathbb{R}^m$.

$\mathcal{L}(\Omega)$ – the linear space of *bounded* random variables from Ω to $\mathbb{R}$

E_P – an expectation operator with respect to a probability measure P.

$\underline{E}$, $\underline{E}_M$ – a lower expectation and a lower expectation with respect to a set of distributions M, respectively.

$\underline{P}$, $\underline{P}_M$ – a lower probabilility and a lower probability with respect to a set of distributions M, respectively.

ν – when used in the context of the imprecise Dirichlet model (IDM), ν is the apriori fixed concentration parameter of the prior set of Dirichlet distributions of the IDM, $\{\text{Dirichlet}(\nu\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \overline{\triangle^p}\}$ (for a fixed, finite p number of categories).

Chapter 1

Elicitation as a statistical motivation for imprecise models

Throughout this thesis, we distinguish between the following.

- By a *precise model,* we refer to the usual Bayesian set up with a single prior and a single likelihood producing a single posterior distribution. Note that hierarchical and mixture models with a single hyper-distribution at the top of the hierarchy are precise.
- By an *imprecise model*, we mean a statistical model consisting of a *set* of Bayesian prior distributions and a single likelihood model that are used to obtain a corresponding set of posterior distributions, one for each prior.

A major theme in this thesis is that imprecise models can be a natural and intuitive statistical tool due to the fact that (prior) elicitation does not always result in a single distribution, but rather a collection that cannot be further whittled down with the information at hand.

1.1 A set of models as a more representative elicitation

We illustrate the need for such models with the following examples.

Example 1.1.1: (Elicitation of a finite number of moments) During elicitation, typically only a finite number of moments can be elicited from an expert (O'Hagan [62]). However, specifying these may not identify a single distribution. For example, Lindsay and Basak [54] observes that matching the first $2p$ moments of a distribution F to a standard normal distribution results in large values of deviations $|F(x) - \Phi(x)|$ in the non-tail region where $|x|$ is small. For example, they report that when F is matched to the first $2p = 60$ moments, $|F(0) - \Phi(0)| \leq 0.2233$ meaning that moment matching does not guarantee a tight fit between F and Φ at $x = 0$. Thus, a finite number of elicited moments cannot be guaranteed to specify a single distribution, and the elicited information would result in a set of distributions instead in these cases.

■

Example 1.1.2: (Elicitation of population parameters in an interval) Typically, most parameters cannot be elicited to an arbitrary degree of precision. This may be due to the following reasons.

- Limits of communication: The ability of the expert to articulate and communicate to the statistician as well as the ability of the statistician to comprehend and 'recover' the original meaning of the information communicated typically determine how accurate the expert information is translated into statistical quantities. This problem is explored in more detail in O'Hagan et al [63].
- Limits of the expert knowledge: communication issues aside, the expert may only be able to specify a statistical quantity up to an interval precision.

Due to the finite precision of the elicited parameters, multiple candidate distributions may be identified as a result.

■

Example 1.1.3: (Combining experts' opinion in isolation) It is sometimes desirable to synthesise a single prior model from the opinions of two or more experts. Garthwaite,

Kadane and O'Hagan [39] provide a comprehensive account of eliciting a single prior probability distribution in this case. We outline two considerations that present obstacles on the path towards eliciting a single prior distribution.

Pooling methods: When individual priors are elicited from experts in isolation, two so-called opinion pools may be formed. The *linear opinion pool* is a convex mixture of the individual elicited prior distributions and the *logarithmic opinion pool* is their weighted geometric mean with weights summing to unity. Which pooling method should be used? Garthwaite, Kadane and O'Hagan, for example, states that the linear pooling satisfies a consistency in marginalisation whereas the logarithmic pooling satisfies the Bayesian externality criteria, meaning that the pooling of the posterior distributions should coincide with the posterior distribution computed from pooling the priors (Madansky [55]). The point is that both properties are considered statistically desirable, and yet each of these two common pooling methods satisfy only one of them [39]. It is not straightforward to choose which pooling method to use unless one has further prior information.

Weights determination: With a choice of the pooling method, the weights to each expert need to be assigned values. Two considerations come to mind: what definition or meaning do the weights have, and how can one verify that the resulting criteria are satisfied?

Let us illustrate these issues with an example of such weight assignments. Garthwaite, Kadane and O'Hagan [39] remark that some experts may be less informed than others and the prior should reflect the difference in credibilities. For example, O'Hagan [62] notes that Cooke [22] proposed to assign weights to each single prior distribution commensurate to each expert's credibility. This results in a prior that is a single mixture or convex combination of the priors with the weights being a hyper-prior distributions on the elicited prior distributions. The weights themselves are determined by having each expert answer relevant questions about the field, the answers to which the statistician knows but the expert does not. In this way, the expert's credibility is checked against a benchmark.

One major issue of this methodology is that the quality of the benchmark is dependent on the statistician's competency with the expert domain. The statistician may err on the side of caution and use more elementary questions that can be easily verified by the statistician. The questions may be so simple that all the experts will pass the benchmark, such that it does not have much power to discriminate the experts' competencies. On the other hand, using more ambitious questions may result in the statistician's inability to verify the

answers, thus invalidating the benchmark itself.

At the other extreme, the ideal benchmark will allow the statistician to exactly compute weights which are commensurate to the benchmark scoring and thus identify a single set of weights. It is likely that reality falls in between, resulting in a non-singleton subset of weights that are candidates to reflecting the experts' competencies.

■

Example 1.1.4: (Combining experts' opinion in group) A second case of combining experts' opinions is when a prior is elicited when the experts act as a group, as opposed to combining the priors elicited from each individual expert. Here, the experts are to reach, as a group that can directly interact with one another, a consensus to the questions posed during elicitation. This is a *behavioural aggregation* method of group elicitation [39]. This method naturally introduces biases due to the psychological effects of working in a group (such as the possibility of dominating personalities affecting and suppressing the views of others or that a final consensus is not possible for other reasons [39]). To partially mitigate these effects, Garthwaite, Kadane and O'Hagan [39] considered also the *Delphi method* whereby each expert's beliefs are separately elicited, aggregated and then shared to all other experts in individual isolation, and these steps are iterated.

Garthwaite, Kadane and O'Hagan [39] also propose two rational ways of reaching a consensus. The first is through a majority voting of ranking of all alternatives (such as the stochastic preference of indicator variables when eliciting probabilities). This is possible only in the most trivial of cases: Arrow's theorem (Arrow [5]) implies that no single set of such preferences can represent a group consensus (that satisfies some conditions for good mathematical behaviour) when there are at least three alternatives to be considered with non-dictatorial rational agents.

The second way discussed by Garthwaite, Kadane and O'Hagan [39] is to consider a single Bayes' model representing the group belief that satisfies a so-called *(weak) Pareto principle*: that is, this group prior preserves all stochastic preferences agreed upon by the individuals of the group. In only a few cases does a prior distribution representing such a set of preferences exist (Garthwaite, Kadane and O'Hagan cite Seidenfeld, Kadane and Schervish [72] and Goodman [42] for such existence conditions).

In conclusion, as Garthwaite, Kadane and O'Hagan [39] have observed: group dynamics prevent a definitive assertion and measurement of consensus. Without consensus, it is unclear what information from an interacting group of experts a single prior distribution actually represents. A *set* of prior distributions is a more natural representation in this situation.

■

Example 1.1.5: (Bounds on quantiles specify a set of distribution from a Beta family) Quantiles are sometimes considered psychologically easier to elicit from an expert as they are easier to comprehend than moments (O'Hagan [62]). However, we encounter the same problem as before, where a finite number of quantiles may only identify a set of distributions.

Suppose that it has been elicited that a candidate prior for a Bernoulli probability θ should be in the beta family of distributions,

$$\left\{ B_{a,b}(p) = \int_0^p \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} t^{\alpha-1}(1-t)^{b-1} dt : \ a, b > 0 \right\}.$$

We are interested in the situation when lower and upper bounds of the 25-th and 75-th quantiles have also been elicited. The α-th quantile of a random variable θ following a beta distribution with parameters a, b is given by,

$$\inf \{t : \ B_{a,b}(t) \geq \alpha\}.$$

In the left panel of Figure 1.1, the level curves of the 25-th and 75-th quantiles,

$$a, b \longmapsto \inf \{t : \ B_{a,b}(t) \geq \alpha\},$$

for $\alpha = 0.25$ and $\alpha = 0.75$ are respectively plotted in red and blue over the set of $\mathbb{R}^+ \times \mathbb{R}^+ \ni (a, b)$.

Now suppose that the elicited bounds are,

$$0.2 \leq \inf \{t : \ B_{a,b}(t) \geq 0.25\} \leq 0.3,$$

and,

$$0.7 \leq \inf \{t : \ B_{a,b}(t) \geq 0.75\} \leq 0.8.$$

The level curves that satisfy these restrictions are plotted (in higher density) in the right panel of Figure 1.1. Importantly, note that the region of intersection of the red and blue curves form the natural parameters of the set of distribution that represents exactly the elicited information: if no other information is given, there is no motivation to prefer one prior over another within this set. In other words, further elicited information or assumptions are needed to distinguish a single prior out of this set as being more suitable than the rest.

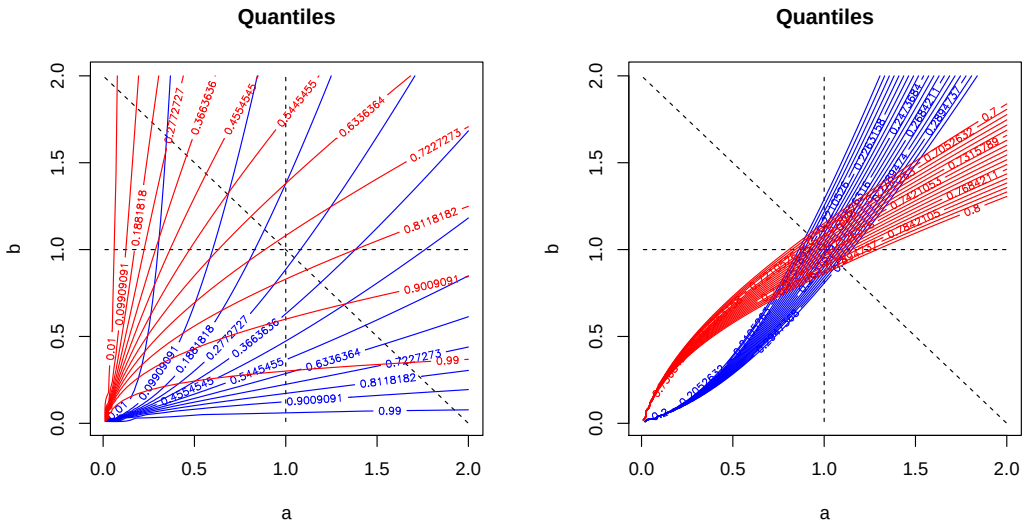

Figure 1.1: Left: the Beta 25-th (red) and 75-th (blue) quantile functions level curves as a function of its hyperparameters, a, b. Right: the same level curves, but restricting the quantile levels to the 25-th quantile being in $[0.2, 0.3]$ (red) and the 75-th quantile functions being in $[0.7, 0.8]$ (blue).

■

1.2 Sensitivity analysis and sets of distributions

As we have seen, it is quite common that the elicited information does not distinguish a single prior distribution choice, but merely leaves one with a set of candidate distributions. In these situations, additional assumptions not implied by the elicitation must be used in

order to pick out a single distribution. If one does choose a single prior out of this set and computes the posterior inference with it, then a *sensitivity analysis* is typically prescribed to check if the posterior inference is sensitive to the additional assumptions made to make this choice. See section 4.7 of Berger [11] for an extensive review of this topic.

A complete sensitivity analysis seeks to understand how inference using the posterior distribution,

$$P(\theta \in \cdot | x) = \frac{\int_{\cdot} L(x|\theta) P(d\theta)}{\int_{\Theta} L(x|\theta) P(d\theta)},$$

can change when the components L, the likelihood model, $P(\cdot)$, the prior model and x, the observed data change. For example, the frameworks of Zhu and Ibrahim [88] and Clarke and Gustafson [21] measure changes of posterior quantities on the left with respect to changes in all three components. However, a sensitivity analysis more commonly refers to the sensitivity towards the prior specification (for example, Berger [10], Gustafson [43], McCulloch [60], Ruggeri and Sivaganesan [68]). In this thesis, we follow the second approach and make the prior model the main focus out of the three components.

1.2.1 Global sensitivity analysis: Prior sensitivity analysis can be categorised as *global* and *local* sensitivity analyses, and this distinction will become relevant to motivating the imprecise methodology from a statistical perspective. 'Global' is meant in the sense that the change of posterior inference is taken over the entire prior model space. For example, if $\mathcal{P}$ denotes all the candidate models from which a single prior model on $\Theta \ni \theta$ is to be chosen, Ruggeri and Sivaganesan [68] cite that the *range (of the posterior expectation of a statistic $T(\theta)$)*,

$$\sup_{P \in \mathcal{P}} E_P(T(\theta)|x) - \inf_{P \in \mathcal{P}} E_P(T(\theta)|x),$$

(where,

$$E_P(T(\theta)|x) = \frac{\int T(\theta) L(x|\theta) P(d\theta)}{\int L(x|\theta) P(d\theta)},$$

for a fixed set of observations x), is a popular metric for sensitivity across all candidate models, whose properties are detailed by, for example, in the overview by Berger et al. [9]. Importantly, for a fixed statistic of interest, the larger the range, the greater the variation of the posterior expectation over the prior model space. Importantly, this is not to be confused with the variation due to the randomness inside of a single prior as measured, for example, by the posterior variance of T (although Ruggeri and Sivaganesan [68] make a case to scale the range with a variance to obtain a more comprehensive picture of the

variations due to both of these factors).

Example 1.2.1: (Global sensitivity analysis over Beta family with quantile restrictions) Continuing with Example 1.1.5, suppose that one wishes to perform a global sensitivity analysis after computing the posterior expectation (conditional on observing the number of successes) of the Bernoulli probability θ with a prior chosen from the set of beta priors with the additional quantile restrictions given in the example. The range introduced by Ruggeri and Sivaganesan [68] in this example is the difference between the maximum and minimum of the posterior expectation as a function over the set of parameters beta a, b whose beta distribution also satisfies the quantile restrictions.

This optimisation may be done in terms of both the natural and mean parametrisation of the Beta family. To this end, in Figure 1.2, we plot the domain of optimisation in the natural parameter and the same domain, but in the mean parameters,

$$\mu_1(a,b) = \gamma^{(1)}(a) - \gamma^{(1)}(a+b),\ \mu_2(a,b) = \gamma^{(1)}(b) - \gamma^{(1)}(a+b),$$

where $\gamma^{(1)}$ is the digamma function.

Notice that in neither parametrisations is the domain convex. This generally does not guarantee that tools of convex optimisation problems may be applied, and makes the optimisation overall challenging without such tools. (We explore a special case of this optimisation problem in Chapter 5.)

■

1.2.2 Local sensitivity analysis: On the other hand, 'local' sensitivity analysis focuses on the variation of posterior quantities in a neighbourhood of the chosen prior model. Note that it is 'the' chosen model in the sense that a single model is first chosen and inference is performed with it. This makes the sensitivity analysis a post-hoc diagnostic and not part of the inference itself.

It turns out that many of the methodologies of local sensitivity analysis share the same idea. If the posterior is treated as a mapping from a prior model to a real number, then how 'flat' is this surface at the point of the chosen prior? The flatter it is, naturally the less the posterior expectation varies over a set around it. The different methods then boil down to how this variation is measured and how this neighbourhood is defined, and we list

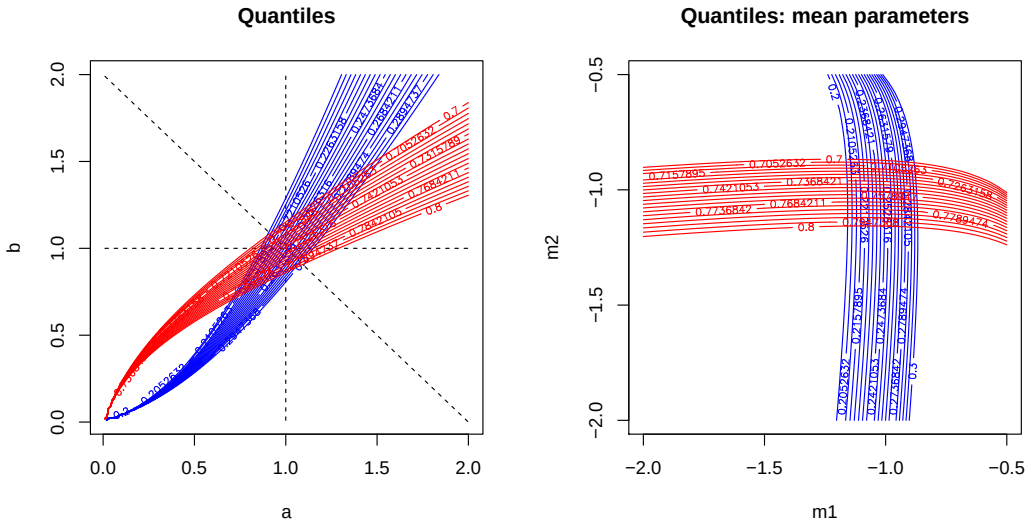

Figure 1.2: Left: the Beta 25-th (red) and 75-th (blue) quantile functions level curves as a function of its natural parameters, a, b restricted to the quantile levels to the 25-th quantile being in $[0.2, 0.3]$ (red) and the 75-th quantile functions being in $[0.7, 0.8]$ (blue). Right: same level curves over mean parametrisation.

a sample of such methods. McCulloch [60] measures sensitivity using the Kullback-Leibler divergences: if the second derivative of the divergence around the chosen prior model is low but the divergence around the resulting posterior model is high, then it means that small changes of prior is associated with large changes in the posterior. Diaconis and Freedman [35] and Ruggeri and Wasserman [69] define a Fréchet derivative of the posterior expectation of statistic as a function of the prior model and use the magnitude of the derivative operator evaluated at the chosen prior model as a measure of how sensitive the change of the posterior expectation is to the prior model. Methods such as those due to Kurtek and Bharath [51] Zhu and Ibrahim [88] formulate similar calculations in a more explicitly geometrical manner that leads to using the differential manifold structure to define the neighbourhood. Lastly, methods such as those due to Marriott and Maroufy [57] use the concept of local mixing (for example Marriott [59]) to construct a neighbourhood that is convex and linear, leading to more tractable computations.

1.2.3 Global or local?: We conclude this review of sensitivity analysis by contrasting global and local analyses. If computationally tractable, a global analysis would by definition yield more information about sensitivity than local. However, one might prefer the local analyses over the global exactly because the latter is computationally too expensive or intractable. In particular, one may be only interested in certain types of perturbations (for example an ϵ contamination neighborhood over a collection of distributions (for example, see Berger et al. [12]). Finally, a local sensitivity implies global sensitivity, such that one might wish to check the former as a sufficient condition when it is straightforward to test first.

1.3 Imprecise methodology and sets of distributions

The *imprecise* methodology we study in this thesis is closely related to the global sensitivity analysis. We will use our previous discussion about the latter to introduce the former in the statistical context.

As we will see in the later chapters, the models in the imprecise methodology which we will be exploring can be largely represented by a set of models, say M. Similar to the sensitivity analysis methodology, we will be working with a set of priors that each produces a posterior distribution, and we will be computing ranges of such sets of posterior expectations, as

well as the end points of the resulting intervals of the form

$$\left[\inf_{P\in M} E_P(T(\theta)|x), \sup_{P\in M} E_P(T(\theta)|x)\right].$$

In this methodology, the length of this interval is called *imprecision*.

1.3.1 Sensitivity analysis versus imprecise methodology: If the imprecise methodology is so similar to that of the global sensitivity analysis, why do we consider the former?

Firstly, the interpretations of the set of distributions in the imprecise setting and the global sensitivity analysis setting are different. In the former, the set of priors is the object being assessed as a model. For example, one might form a set of such prior distributions by translating elicited information into restrictions over a space of distributions.

Example 1.3.1: Consider $\Theta = \{\theta_1, \theta_2, \theta_3\}$ to be the space of possible likelihood parameters in consideration. A prior on this space is specified by $P(\{\theta_i\}) = p_i$, for the triplet $\boldsymbol{p} = (p_1, p_2, p_3) \in \triangle^3$ in the unit simplex of $\mathbb{R}^3$. Let $T(\theta)$ be a statistic over the model space. One might elicit from an expert bounds on the expectation of a finite number of random variables $f_1(T), \ldots, f_p(T)$, each in the form of,

$$f_i(T(\theta_1))p_1 + f_i(T(\theta_2))p_2 + f_i(T(\theta_3))p_3 \leq c_i,$$

where c_i is the elicited upper bound of the expectation of f_i under $\boldsymbol{p}$. In other words, the elicited information about the prior model is that the prior should satisfy,

$$\boldsymbol{F}\boldsymbol{p} \leq \boldsymbol{c},$$

where $\boldsymbol{F} = [f_i(T(\theta_j))]$ is a $p \times 3$ matrix and $\boldsymbol{c} = [c_i]^T$ is a $p \times 1$ vector of real numbers. When feasible, this system represents a convex polytope in $\triangle^3$. That is to say, the distributions in this *set of priors* are all consistent with the elicited information, and no single one is preferred over another in this set. Thus, at this apriori stage, this entire set should be considered 'a prior model', as opposed to a single prior in this set. In the imprecise methodology, the entire set is to be used in posterior inference and in the subsequent chapters of this thesis we will review how this is done in a *coherent* manner in the same way that Bayesian probabilities are coherent (see Lindley [53], Jeffrey [49] or Definition 2.1.2, for example, for a definition of coherence for Bayesian probabilities).

■

In contrast, the set in a sensitivity analysis is not elicited in the same sense as the prior distribution of which it is a neighborhood. Rather, it is a post hoc construction specifically for testing the robustness of the inference, and is not part of the inference itself.

This leads to different interpretations between imprecision and range, even as they measure the same quantity over the respective sets of distributions. However, because the set is part of the model in the imprecise methodology, it is not to be interpreted as a measure of robustness in the usual sense of the range. Rather, the *imprecision* is to be taken directly as one of the posterior statistics to be reported as part of the inference, at the same level as say posterior means, quantiles and variances. In fact, we draw the following analogy: just as in Bayesian inference where the posterior distribution is considered to embody the inference itself (with a sensitivity analysis being post-hoc and considered separate), the set of posterior distributions resulting from the elicited set of prior distributions is also to be considered to embody the inference and not to be treated as a diagnostic tool.

One issue of working with a set of distributions as a model is that Bayesian inference typically works with coherent probabilities: the coherence of a *set of probabilities* is typically not defined in common settings. (See Chapter 2.) Another reason why sensitivity analysis is considered post-hoc and not part of the inference is that the sensitivity analysis methodology does not have a set of rules that defines it to be coherent, unlike (Bayesian) probabilities (again see Definition 2.1.2). This fact highlights a difference between a (global) sensitivity analysis and an imprecise methodology: there is a definition for coherence of a set of models in the latter, which is why it can be readily claimed to produce principled and coherent inference directly using a set of models as opposed to just one.

Example 1.3.2: We will see that a form of the *generalised Bayes' rule* (GBR) (see Theorem 2.2.4) in the imprecise methodology provides a coherent manner of constructing posterior inference from a set of prior distributions. To continue with our earlier example, if we let $\theta \longmapsto L(x|\theta)$ to be a fixed likelihood model, then a typical way to obtain an imprecise posterior inference is to form a set of posterior expectations,

$$M_{|x} = \left\{ U \longmapsto \frac{\sum_{i=1}^{3} U(\theta_i) L(x|\theta_i) p_i}{\sum_{i=1}^{3} L(x|\theta_i) p_i} \; : \; \boldsymbol{p} \in \triangle^3 \; \wedge \; \boldsymbol{F}\boldsymbol{p} \leq \boldsymbol{c}, \; \sum_{i=1}^{3} L(x|\theta_i) p_i > 0 \right\},$$

(where x is some observation from the likelihood model). A typical inference that might

be reported about $T(\theta)$ is,

$$\left[\inf_{P \in M_{|x}} E_P(T(\theta)|x), \sup_{P \in M_{|x}} E_P(T(\theta)|x)\right].$$

The real numbers in this interval are then interpreted as values which are consistent with the data, the likelihood and the set of priors that generated the posterior inference.

■

Thus, we have another motivation to consider the imprecise methodology. Coherence allows for logical consistency when working with Bayesian probabilities (Lindley [53]), but it needs to be defined for a set of prior distributions. Without such an extension, a Bayesian has to perform inference with a single prior and essentially work with a candidate set through the external and post-hoc methodology of sensitivity analysis. From our earlier discussion, we have concluded that sometimes a single prior distribution might not be deducible from the elicited information alone. In this light, the imprecise methodology allows one to work with a set of prior models apriori while maintaining coherence.

Chapter 2

Review of aspects of imprecise models

In this chapter, we critically review Walley's [81] *theory of imprecise probabilities*, the theory upon which this thesis is based. In addition to being a literature review, we will also provide our own pedagogical examples to illustrate the concepts in the theory for a statistical audience.

Throughout this thesis, we use *imprecise probability* as a tool to construct posterior inference. By considering *imprecision* as part of the model as opposed to a separate indicator of reliability of the model, posterior inference from *imprecise models* does not force the choice of a single prior. This yields a model for statistical inference that is capable of simultaneously taking multiple prior models into consideration.

As we have alluded in Section 1.3, *imprecise expectations*, which are a procedure of taking the minimum and maximum of expectations over a set of distributions, is methodologically different from a sensitivity analysis over a set of distributions. Sensitivity analysis does not prescribe principled approaches to either picking a single distribution from a set of candidates or working with the whole set. Walley's theory of imprecise probabilities, the theory behind imprecise expectations, in addition to its mathematical component, also prescribes how to construct reasonable models, in the form of the concepts of *avoiding sure losses* and *coherence*.

2.1 Avoiding sure losses and coherence

Let us first give a definition of *assessment* to pin down our usage of this term throughout this thesis.

Definition 2.1.1: An *assessment* is an assignment of (imprecise or precise) expectation values to a set of random variables.

□

Example 2.1.1: For sets A and B, there are at least two ways to provide an assessment of the probability $P(A \cup B)$. One is use the inclusion-exclusion principle $P(A) + P(B) - P(A \cap B)$ if one already has these probabilities at hand. Another is to consider $C = A \cup B$ directly: this can be done, for example, by eliciting information about $P(C)$ from an expert. (Notice that the expert need not know that C is in fact the union of A and B.)

Importantly, this example distinguishes an assessment which is derived from (and therefore obeys) the laws of probability from one whose value may not do so. The latter may potentially contradict other existing assessments of probabilities.

■

The two main principles driving the axioms of imprecise probabilities are *avoiding sure losses* (assessments of probabilities that incur such losses are called *Dutch books*) and *coherence*. Importantly, sensitivity analysis does not implement these concepts, causing it to diverge from the methodology of imprecise probability models. We review how coherence and avoiding losses are implemented in the imprecise probability theory, which will be the workhorse of this thesis.

We begin with the following definition from Lindley [53].

Definition 2.1.2: (Section 5.4 of Lindley [53]) For a sample space Ω with subsets $A, B \subset \Omega$, an assessment $P(\cdot), P(\cdot|\cdot)$ over the subsets of this space is *coherent* iff it satisfies,

$$\begin{gathered} P(A), P(B) \in [0,1], \\ P(\Omega) = 1, \\ P(A \cup B) = P(A) + P(B), \text{ when } A \text{ and } B \text{ are disjoint}, \\ P(A \cap B) = P(A)P(B|A). \end{gathered}$$

□

Authors of foundational probability theory such as de Finetti [32], Savage [70], and Lindley [52], [53] have motivated this definition by a so-called *subjective Bayesian* viewpoint (Walley [81]). It provides qualitative requirements for consistency of the assessments via a *gambling analogy*. A linear utility is established to measure a gambler's gain or loss. A gambler engages in a *gamble* whose random reward is the realisation of a random variable. Elicitation and assessment of probabilities of events and expectations of random variables are treated as assigning prices to such gambles which are acceptable for the gambler given certain knowledge about the realisation of the generating process. In principle, the gambler should not accept prices that lead to a systematic losses and prices should be internally consistent. *Avoiding sure losses* and *coherence* necessarily avoid two main types of such inconsistencies in probability assessments.

2.1.1 Probability distribution and expectation

It is more useful to frame these inconsistencies in terms of expectations. In this context, let us understand first what avoiding sure losses precisely entails. First, we follow Walley [81] and make the following assumption.

Definition 2.1.3: For a sample space Ω, let $\mathcal{L}(\Omega)$ be the linear space of bounded random variables over Ω. (The space is linear because finite sums of bounded random variables is again a bounded random variable.)

□

Condition 2.1.1: Unless stated otherwise, all random variables are *bounded* and in $\mathcal{L}(\Omega)$[1].

□

[1]This is in accordance with Walley [81] whom we follow closely. For discussions of imprecision involving unbounded random variables, see Sections 3.3.1 in this thesis. For an in-depth treatment of extensions to extended-real-valued random variables of the theory of imprecise probabilities, see Troffaes and de Cooman [78].

The following is a special case of the more general definition stated in Walley [81], given in Definition 2.1.7.

Definition 2.1.4: Given a sample space Ω, an assessment of expectation E_P *incurs sure losses* (or is a *Dutch book*) iff it is an assessment over a set of random variables $\mathcal{F} \subseteq \mathcal{L}(\Omega)$, $E_P : \mathcal{F} \longmapsto \mathbb{R}$, such that there exists random variables $X_1, \ldots, X_n \in \mathcal{F}$ such that:

$$\forall \omega \in \Omega : \sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) < 0. \tag{2.1}$$

An assessment that does not incur sure losses is said to *avoid sure losses*.

□

An expectation that avoids sure losses ensures that we do not have any finite combination of zero-expectation random variables being pointwise negative. Otherwise, it contradicts the principle that such a sum should itself have a zero expectation. Expectations of a single probability distribution typically avoid sure losses.

Lemma 2.1.1: (Lemma A.2.1) For any sample space Ω, and P a distribution over some σ-field of Ω, any expectation E_P over all bounded random variables avoids sure losses. In other words, for any $X_1, \ldots, X_n$ that are bounded,

$$\sup_{\omega} \sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) \geq 0.$$

□

Despite being seemingly trivial in the probabilistic setting, we will see in Example 2.1.2 that avoiding sure losses is not at all trivial when attempting to ensure lack of losses over multiple distributions, and we have seen from Chapter 1 that the latter could occur commonly in statistical practice.

A less severe inconsistency amongst assessments is *incoherence*. The following is a special case of Definition 2.1.8, which represents a more general definition by Walley [81].

Definition 2.1.5: Given a sample space Ω, an assessment of expectation E_P is *incoherent* if it is an assessment over a set of random variables $\mathcal{F} \subseteq \mathcal{L}(\Omega)$, $E_P : \mathcal{F} \longmapsto \mathbb{R}$, such that

there exists random variables $X_0, X_1, \ldots, X_n \in \mathcal{F}$ and $m \in \mathbb{N}$ such that:

$$\forall \omega \in \Omega : \sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) < m(X_0(\omega) - E_P(X_0)). \tag{2.2}$$

An assessment that is not incoherent is *coherent*.

□

Let us qualitatively unwrap this definition. The random variables $\sum_i (X_i - E_P(X_i))$ and $X_0 - E_P(X_0)$ both have an expectation of zero. What (2.2) means is that a sum of random variables with a zero expectations is strictly less than any other positively scaling of any other random variable with a zero expectation, implying in contradiction that one of the sides does not have a zero expectation.

Importantly, this is less severe than avoiding sure losses as the assessments for the expectation of a random variable can still be bounded between the latter's infimum and supremum, so they are individually consistent with each other. Indeed, if the assessment $E_P(X_0)$ is *incoherent* with the rest in (2.2), it is intuitively clear that $E_P(X_0)$ can be corrected by increasing it so as to bring the right side of the inequality closer to the left[2].

Again, expectations of random variables from a single distribution are coherent.

Lemma 2.1.2: (Lemma A.2.2) For any sample space Ω, and P a distribution over Ω over some σ-field of Ω, any expectation E_P over all bounded random variables are coherent. In other words, for any $X_0, X_1, \ldots, X_n$ that are bounded and $m \in \mathbb{N}$,

$$\sup_{\omega} \sum_{i=1}^{n} ((X_i(\omega) - E_P(X_i)) - m(X_0(\omega) - E_P(X_0)) \geq 0.$$

□

Like avoiding sure losses, we will see in Section 2.1.2 that coherence is not trivial when considering multiple distributions simultaneously.

[2]The minimum value to which to increase the assessment is called a *natural extension* of $E_P(Z)$ (relative to the rest of the assessments).

2.1.2 Sets of probability distributions and expectations

Despite the fact that probability assessments following Definition 2.1.2 avoids sure losses and are coherent, such losses and incoherence may be unnoticeably present in practice when considering sets of distributions. Throughout this thesis, we follow Walley [81] and focus on the *lower and upper (or imprecise) expectations* as a summary over a set of distributions.

Condition 2.1.2: Whenever given a set of distributions M over which we compute the expectation of a random variable $X \in \mathcal{L}(\Omega)$, we assume that it is measurable against an existing σ-field shared by all distributions in M: we will simply say that X is *suitably measurable* in this case.

□

Definition 2.1.6: Let M be a set of distributions over some sample space Ω with a suitably chosen σ-field. Define the *lower and upper expectations*[3] and similarly of a suitably measurable random variable $X \in \mathcal{L}(\Omega)$, as,

$$\underline{E}_M(X) = \inf \{E_P(X) : P \in M\},$$

and

$$\overline{E}_M(X) = \sup \{E_P(X) : P \in M\},$$

respectively. We will loosely refer to one or the pair of lower and upper expectations as *imprecise expectations*.

□

Notice that, $\overline{E}_M(X) = -\underline{E}_M(-X)$ such that one can compute the upper expectation given the lower expectation[4]. As a result, where appropriate, we focus our analyses on the lower expectation.

Definition 2.1.7: (Walley [81], 2.4.1) Given a sample space Ω, an assessment of lower expectation $\underline{E}_M$ from a set of probability distributions M *incurs sure losses* (or is a *Dutch*

[3]For reason of clarity, we forgo the typical use of $\underline{E}$ and $\overline{E}$ as natural extensions of lower and upper previsions in the imprecise probabilities literature.

[4]$\underline{E}_M$ and $\overline{E}_M$ are said to be a *conjugate* pair of lower and upper previsions (Walley [81]).

book) iff it is an assessment over a set of random variables $\mathcal{F} \subseteq \mathcal{L}(\Omega)$, $\underline{E}_M : \mathcal{F} \longmapsto \mathbb{R}$ such that there exists random variables $X_1, \ldots, X_n \in \mathcal{F}$ such that:

$$\forall \omega \in \Omega : \sum_{i=1}^{n} (X_i(\omega) - \underline{E}_M(X_i)) < 0. \tag{2.3}$$

An assessment that does not incur sure losses is said to *avoid sure losses*.

□

Definition 2.1.8: (Walley [81], 2.5.1) Given a sample space Ω, an assessment of lower expectation $\underline{E}_M$ from a set of probability distributions M is *incoherent* if it is an assessment over a set of random variables $\mathcal{F} \subseteq \mathcal{L}(\Omega)$, $\underline{E}_M : \mathcal{F} \longmapsto \mathbb{R}$ such that there exists random variables $X_0, X_1, \ldots, X_n \in \mathcal{F}$ and $m \in \mathbb{N}$ such that:

$$\forall \omega \in \Omega : \sum_{i=1}^{n} (X_i(\omega) - \underline{E}_M(X_i)) < m(X_0(\omega) - \underline{E}_M(X_0)). \tag{2.4}$$

An assessment that is not incoherent is *coherent*.

□

We note here that these definitions are not new, even at the time Walley [81] was written. Indeed, Walley noted that these definitions were previously investigated by Huber [48], Smith [74] and Williams [86] [85]. For later chapters, we also define *imprecise probabilties* over a set of distribution.

Definition 2.1.9: For a set of distributions M, define the *lower and upper probabilities* of a suitably measurable event A to be,

$$\underline{P}_M(A) := \underline{E}_M(I_A), \quad \overline{P}_M(A) := \overline{E}_M(I_A) \ ,$$

where I_A is the indicator function of the event A. We say that $\underline{P}$ and $\overline{P}$ are coherent iff $\underline{E}$ and $\overline{E}$ are respectively coherent.

□

We illustrate sure losses by considering the following case involving two random variables, whose individual assessments on their expectations avoids sure losses but a sure loss can be generated when they are considered jointly.

Example 2.1.2: Let the sample space consist of three categories, and

$$X_1 = (-2, 0, -1), \quad X_2 = (0, 1, 2), \quad Y_1 = (-1, 0, 1), \quad Y_2 = (0, 1, 0),$$

be random variables on this sample space. Consider an assessment about these random variables by combining the elicitations from two different experts. The first expert has marginal information about X_1 and X_2 in the form of lower bounds for their expectations that reflect the elicited information. That is, a probability distribution P of this system should generate expectations E_P satisfying,

$$E_P(X_1) \geq -0.05, \quad E_P(X_2) \geq 1.75 \ .$$

The second expert has information about Y_1 and Y_2 in the form of upper and lower bounds for their respective expectations, such that,

$$E_P(Y_1) \leq 0, \quad E_P(Y_2) \geq 0.9 \ .$$

Consider that,

$$\sup_{\omega \in \{\omega_1, \omega_2, \omega_3\}} (X_1(\omega) + X_2(\omega) - Y_1(\omega) + Y_2(\omega)) - (-0.05 + 1.75 - 0 + 0.9) = 2 - 2.6 = -0.6.$$

Why is this problematic? Write M to represent a subset of probability distributions over Ω that satisfies the elicited bounds, and,

$$\underline{E}_M(X_1) = -0.05, \quad \underline{E}_M(X_2) = 1.75, \quad \overline{E}_M(Y_1) = 0, \quad \underline{E}_M(Y_2) = 0.9,$$

to represent the bounds. Then,

$$\begin{aligned}
&\sup_{\omega \in \{\omega_1, \ldots, \omega_3\}} (X_1(\omega) + X_2(\omega) - Y_1(\omega) + Y_2(\omega)) \\
&< \underline{E}_M(X_1) + \underline{E}_M(X_2) - \overline{E}_M(Y_1) + \underline{E}_M(Y_2) \\
&= \inf_{P \in M} \{E_P(X_1)\} + \inf_{P \in M} \{E_P(X_2)\} - \sup_{P \in M} \{E_P(Y_1)\} + \inf_{P \in M} \{E_P(Y_2)\} \\
&= \inf_{P \in M} \{E_P(X_1)\} + \inf_{P \in M} \{E_P(X_2)\} + \inf_{P \in M} \{E_P(-Y_1)\} + \inf_{P \in M} \{E_P(Y_2)\} \\
&\leq \inf_{P \in M} \{E_P(X_1 + X_2 - Y_1 + Y_2)\}.
\end{aligned}$$

In other words, any expectation due to any probability distribution adhering to the assessed bounds results in the expectation of $X_1 + X_2 - Y_1 + Y_2$ being strictly greater than the maximum of this random variable, leading to a contradiction.

■

Let us provide an example where sure losses are avoided by a set of distributions (through their expectations), but they are incoherent.

Example 2.1.3: Consider $X = (-1/8, 1/4)$, $Y = (2, -1)$ and $Z = (3, -1/2)$, a distribution P with the assessments,

$$E_P(X) \geq 0, \quad E_P(Y) \geq 0, \quad E_P(Z) \geq 0 \ .$$

We write the set of distributions satisfying these constraints as,

$$M = \{p \in [0,1] : \ E_p(X) \geq 0 \ \wedge \ E_p(Y) \geq 0 \ \wedge \ E_p(Z) \geq 0\},$$

and the lower expectations,

$$\underline{E}_M(X) = 0, \ \underline{E}_M(Y) = 0, \ \underline{E}_M(Z) = 0.$$

To simplify the algebra, we consider E_P to be restricted to the domain $\mathcal{F} = \{X, Y, Z\}$: this is shown to avoid sure losses over $\mathcal{F}$ in Proposition A.2.1. (For assessments over larger sets of random variables, linear programming is typically used to check avoidance of sure losses. See Quaeghebeur [66] and Walley, Pelessoni and Vicig [82].)

However, consider the random variable $(X - \underline{E}_M(X)) + (Y - \underline{E}_M(Y)) - (Z - \underline{E}_M(Z))$ in the criterion (2.4). By evaluating this point-wise over Ω,

$$\omega = \omega_1 : \ -\frac{1}{8} + 2 - 3 < 0,$$

and

$$\omega = \omega_2 : \ \frac{1}{4} - 1 + \frac{1}{2} < 0,$$

such that this satisfies (2.4) so E_P is incoherent.

Consider again the set of distributions that satisfy the assessed constraints:

$$\begin{aligned}
M &= \{p \in [0,1] : \ E_p(X) \geq 0 \ \wedge \ E_p(Y) \geq 0 \ \wedge \ E_p(Z) \geq 0\} \\
&= \{p \in [0,1] : \ -p/8 + (1-p)/4 \geq 0 \ \wedge \ 2p - (1-p) \geq 0 \ \wedge \ 3p - (1-p)/2 \geq 0\} \\
&= \{p \in [0,1] : \ 2/3 \geq p \ \wedge \ p \geq 1/3 \ \wedge p \geq 1/7\} \\
&= \{p \in [0,1] : \ 2/3 \geq p \ \wedge \ p \geq 1/3\} \\
&= \{p \in [0,1] : \ 2/3 \geq p \geq 1/3\}.
\end{aligned}$$

Notice that the constraint $E_p(Z) \geq 0$ is subsumed into $E_p(Y) \geq 0$. Incoherence, in this case, has to do with inconsistencies due to the assessment for Z: the lower bound of $E_p(Z)$, under the constraints $E_p(X) = 0$ and $E_p(Y) = 0$, is not zero. Indeed, if we consider $E_p(Z)$,

$$E_p(Z) = -\frac{1}{2} + \frac{7}{2}p.$$

Because the constraints $E_p(X) \geq 0$ and $E_p(Y) \geq 0$ is equivalent to $2/3 \geq p \geq 1/3$, the lower bound of $E_p(Z)$ over this set of distributions is in fact,

$$-\frac{1}{2} + \frac{7}{2}\frac{1}{3} = -\frac{3}{6} + \frac{7}{6} = \frac{2}{3} > 0,$$

such that the original assessment $E_p(Z) \geq 0$ is too loose and, therefore, redundant in light of $E_p(X), E_p(Y) \geq 0$.

■

In Example 2.1.3, by removing redundant constraints on M, we have corrected the bounds on $E_p(Z)$ to better reflect the state of information represented by the assessed constraints. In other words, by reviewing a set of assessment that avoids sure losses, we can alter our initial assessments to make them consistent with one another[5].

2.2 Aspects of the theory of imprecise probabilities

2.2.1 Imprecise expectation

We have already touched on this representation in the previous sections where the assessment is a set of distributions satisfying certain constraints over which we optimise to obtain

[5]This alteration is the *natural extension of* $\underline{E}_M(Z)$. See 3.1 of Walley [81].

imprecise expectations. We illustrate the geometry in finite dimensions using subsets of a *probability simplex*, the space of all possible finite dimensional multinomial distributions.

Importantly, we will see that, in this context, coherent sets are *convex sets* and a finite set of moments bounds make this set a *polytope* in this space. We properly review this convex geometry developed in Walley [81] in general cases in Section 2.2.2.

The following finite dimensional geometrical intuitions are useful as guidelines for the case of continuous random variables, and also directly used in Chapter 6.

Example 2.2.1: (Sets of multinomial distributions) Let $|\Omega| < \infty$ and let $\boldsymbol{p} = (p(\omega_i) : \ i = 1, \ldots, |\Omega|)$ denote a probability distribution on Ω. Suppose that, for two bounded random variables Y, Z over Ω, the following bounds on the moments are elicited,

$$\begin{aligned} a \leq Y \cdot \boldsymbol{p} &= \sum_{i=1}^{|\Omega|} Y(\omega_i) p_i \leq b, \\ c \leq Z \cdot \boldsymbol{p} &= \sum_{i=1}^{|\Omega|} Z(\omega_i) p_i \leq d, \end{aligned} \tag{2.5}$$

for real numbers $\min Y \leq a < b \leq \max Y$ and $\min Z \leq c < d \leq \max Z$ and where $\boldsymbol{u} \cdot \boldsymbol{v}$ denotes the dot product between two Euclidean vectors. The assessment of the set of distributions that are consistent with this elicitation is,

$$M = \left\{ \boldsymbol{p} : \ \sum_{i=1}^{|\Omega|} p(\omega_i) = 1, \ a \leq E_{\boldsymbol{p}}(Y), -b \leq E_{\boldsymbol{p}}(-Y), \ c \leq E_{\boldsymbol{p}}(Z), -d \leq E_{\boldsymbol{p}}(-Z), \ \boldsymbol{p}(\omega_i) \geq 0 \right\}.$$

At this point, the assessment corresponding to the elicitation (2.5) is only about Y, Z, $-Y$ and $-Z$. This means that upper bounds on expectations such as $Y \cdot \boldsymbol{p} \leq b$ is simply $\overline{P}(Y) = -\underline{P}(-Y) \leq b$ which yields the restriction $-b \leq -Y \cdot \boldsymbol{p}$ above.

Now, the lower expectation over this assessment is,

$$\underline{E}_M : \ W \ \longmapsto \ \inf \{ W \cdot \boldsymbol{p} : \ \boldsymbol{p} \in M \}.$$

In this finite dimensional space, this is the solution of a linear programming of the objective function $W \longmapsto W \cdot \boldsymbol{p}$ over the closed convex polytope A. Figure 2.1 illustrates some cases

of how this convex polytope may be assessed.

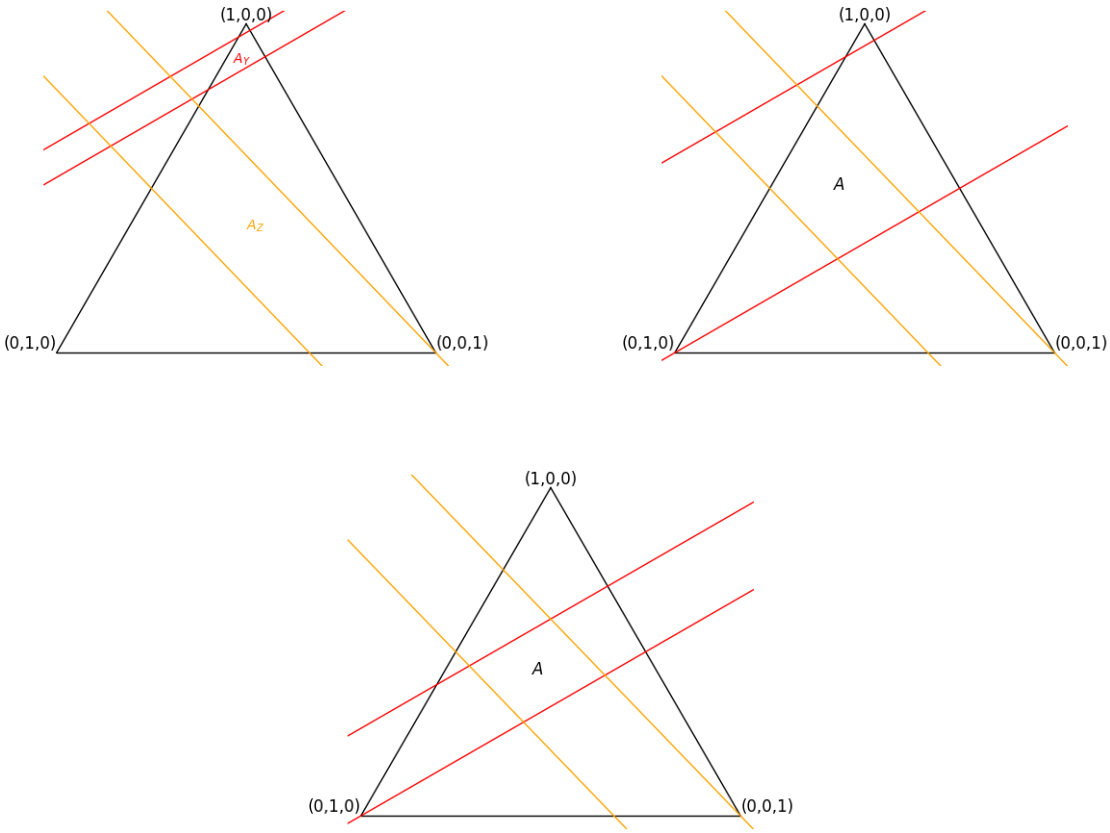

Figure 2.1: Cases for assessments of sets of distributions A of hyperplane boundaries $a \leq Y \cdot \boldsymbol{p} \leq b$ (in red) and $c \leq Z \cdot \boldsymbol{p} \leq d$ (orange). Top left: assessments about Y do not intersect with those about Z such that $A = A_Y \cap A_Z = \emptyset$. Top Right: the assessment $Y \cdot \boldsymbol{p} \leq b$ (top red line) is not used in constructing A, such that $\overline{P}(Y) = b$ does not contribute to A, and is therefore an incoherent assessment. Bottom centre: a coherent assessment.

The top left panel is a case where a, b, c, d are such that the moment conditions are incompatible: that A is in fact empty. In other words, these constraints incur sure losses as bounds on expectations. From an optimisation perspective, the optimisation domain M

contains no feasible solutions. The top right panel is a case where A is in fact not empty, but is defined only by three of the four conditions (and the boundary of the simplex), such that the remaining condition $Y \cdot \boldsymbol{p} \leq b$ is redundant. Thus, the four constraints avoid sure losses but are incoherent due to this redundancy. Finally, the bottom panel depicts an assessment of A where the information from all four assessments are in effect and coherent.

■

We have already covered much of the preliminary construction of lower and upper expectations in Section 2.1.2. Here, we will focus on the key results and concepts of Walley [81] about the imprecise expectations as a result of them being coherent assessments.

2.2.2 Imprecision and vacuity

In terms of imprecise expectations, the notion of *imprecision* represents the variation of the expectation over a set of distributions.

Definition 2.2.1: (Walley [81]) The *imprecision* of X over a set of distributions M is the quantity, $\overline{E}_M(X) := \overline{E}_M(X) - \underline{E}_M(X)$.

□

When the imprecision is maximal for X, we say that the imprecise model is *vacuous* for X, as the set of distribution contains no information about the expectation of X to decrease the imprecision.

Definition 2.2.2: (Walley [81]) A set of distributions M generates *vacuous imprecise expectations for* X iff $\underline{E}_M(X) = \inf_{\omega \in \Omega} X(\omega)$ and $\overline{E}_M(X) = \sup_{\omega \in \Omega} X(\omega)$.

□

We emphasise here that X is in $\mathcal{L}(\Omega)$, the linear space of bounded random variables. When, in addition, $|\Omega| < \infty$, vacuous imprecise expectations of any element X in $\mathcal{L}(\Omega)$ are finite and the imprecision is also finite.

It is interesting that imprecise expectations has a notation for representing noninformativeness through vacuity. This is in contrast with probability models where noninformativeness typically means the use of a uniform distribution as a single model of uncertainty. As Walley argues in [80], a uniform distribution assumes that all states have the same probability of occurring: this itself is a piece of information, which contradicts its purpose of representing noninformativeness.

Example 2.2.2: (Vacuous sets of multinomial models) Using the setting of Example 2.2.1, if instead of the moment conditions, suppose that only Ω has been elicited. Then, the set of distributions consistent with only knowing the finite number of categories of the sample space is simply the set of all possible categorical distributions over Ω, i.e.

$$M = \left\{ \boldsymbol{p} : \sum_{i=1}^{|\Omega|} p(\omega_i) = 1, \ p(\omega_i) \geq 0 \right\} = \overline{\triangle^{\Omega}}.$$

This is the closed unit simplex in $\mathbb{R}^{|\Omega|}$. Notice that, for any bounded random variable X,

$$\underline{E}_M(X) = \min_{\omega \in \Omega} X(\omega), \ \overline{E}_M(X) = \max_{\omega \in \Omega} X(\omega).$$

The optimisers of these solutions are simply the vertices of $\overline{\triangle^{\Omega}}$ where unit probabilities are assigned to the minimum and maximum of X, respectively.

■

2.2.3 The lower envelope theorem

Because $\underline{E}_M$ is a functional that assigns values for the lower bound of expectations for a collection of random variables, we can consider the coherence of such an operator. The following result, called the *lower envelope theorem*, is the main driver of the results in Walley's imprecise probability theory.

Definition 2.2.3: (Walley [81]) The *lower envelope*[6] of a set of distributions M is the functional,

$$\underline{E}_M : \ X \longmapsto \inf \{E_P(X) : \ P \in M\},$$

[6]In the main text of this thesis, we use the notation $\underline{E}_M$ for both the lower envelope of M and lower expectation over M as they often coincide.

over the set of suitably measurable random variables in $\mathcal{L}(\Omega)$.

□

Definition 2.2.4: (Walley [81]) For a functional $\underline{E}$ over $\mathcal{L}(\Omega)$, define,

$$M_{\underline{E}} := \{P \in \mathcal{P} : \ \forall X \in \mathcal{L}(\Omega), \ E_P(X) \geq \underline{E}(X)\},$$

as its *dominating set of distributions*[7]. ($\mathcal{P}$ is the set of all distributions over Ω.)

□

Theorem 2.2.1: (Lower envelope theorem, Walley [81], 3.3.3) A functional $\underline{E}$ is *coherent* iff it is the lower envelope of the set $M_{\underline{E}}$.

□

Furthermore, the lower envelope of a set of distributions is an element of the set.

Theorem 2.2.2: (Extreme value theorem, Walley [81], 3.6.2 (c)) If $\underline{E}$ is coherent, then, for any bounded random variable $X \in \mathcal{L}(\Omega)$, the minimisation,

$$\inf \{E_P(X) : \ P \in M_{\underline{E}}\},$$

is attainable by an element of $M_{\underline{E}}$.

□

We note that Theorems 2.2.1 and 2.2.2 are driven by the fact that, given a lower expectation $\underline{E}$, its dominating set of expectations $M_{\underline{E}}$ is a convex and compact set in the dual space of $\mathcal{L}(\Omega)$. (The convexity is clear from the definition of $M_{\underline{E}}$. See Appendix D of Walley [81] for a discussion about the compactness of $M_{\underline{E}}$.)

[7] In Walley [81], this set is expressed as a set of expectation operators, and is more generally a set of dominating *linear previsions*.

Finally, a coherent lower expectation is bijective to the set of distributions that dominates it.

Theorem 2.2.3: (3.6.1 of Walley [81]) Let $\underline{E}$ be a coherent lower expectation and M be its dominating set such that $\underline{E}$ is the lower envelope of M. Then, the map $\underline{E} \longmapsto M_{\underline{E}}$ and $M \longmapsto \underline{E}_M$ are bijections that form the inverse of each other. In particular,

$$\underline{E}_{M_{\underline{E}}} = \underline{E},$$

and

$$M_{\underline{E}_M} = M.$$

□

2.2.4 Posterior lower/upper expectations

In this section, we introduce a version of the *generalised Bayes' rule* that extends Bayesian posterior calculations to using lower and upper expectations developed in Walley [81]. It guarantees a more general form of coherence called *joint coherence* than Definition 2.1.8, amongst the imprecise prior model, the single likelihood and the resulting imprecise posterior model. To avoid detracting from the review, the relevant points to note are that,

- *joint coherence* (6.3.2 and 7.1 of Walley [81]) is again motivated to avoid Dutch books-like arbitrage amongst multiple conditional and unconditional imprecise models,
- the conditional lower and upper expectations induced by the lower and upper envelopes of a collection of precise conditional distributions due to application of Bayes' rule on each of the unconditional distributions in the assessed set are also jointly coherent with the prior imprecise expectations, and that
- under regularity conditions in Walley [81], any such posterior imprecise expectations are also jointly coherent with the prior imprecise expectations as well as the precise likelihood used.

The so-called *generalised Bayes' rule (GBR)*, Theorem 6.4.1 of Walley [81], is the following equation as a direct consequence of joint coherence,

$$\underline{P}_M(I_B(X - \underline{E}_M(X|B))) = 0. \tag{2.6}$$

It defines the conditional lower expectation $\underline{E}_M(\cdot|B)$ as the solution to the equation that is jointly coherent with the unconditional model $\underline{E}_M$ and generalises the probabilistic Bayes' rule. However, the following lower envelope version of the generalised Bayes' rule is more useful and illustrative of how Bayes' rule is generalised to the imprecise expectations setting.

Proposition 2.2.1: (Passage 6.4.2 of Walley[81]) when $\underline{E}$ is coherent and $\underline{E}(B) > 0$, then the (coherent) conditional lower expectation $\underline{E}(\cdot|B)$ that is the solution to the generalised Baye's rule equation (2.6) is the lower envelope of the pointwise application of the classical Bayes' rule:

$$\underline{E}(X|B) = \min \ \{E_P(XI_B)/E_P(I_B) : \ E_P \geq \underline{E}\},$$

over a domain of suitably measurable random variables.

□

Briefly, this result allows us to achieve joint coherence between our unconditional imprecise model represented by the unconditional lower expectations $\underline{E}$ a set of conditional lower expectations $\{\underline{E}(\cdot|B) : \ B \in \mathcal{B}\}$ where $\mathcal{B}$ is a partition of Ω.

Given data D, for each prior $P \in M$, we construct $P(\cdot|D)$ by applying Bayes' rule to each prior distribution paired with the likelihood function given. With this set of probability distributions of prior distributions, coherent posterior lower and upper expectations of a suitably measurable random variable, say $f(\theta)$, may be had by appealing to Theorem 2.2.4.

Theorem 2.2.4: (Passage 8.4.8 of Walley [81]) Suppose that M is a set of (expectation operators of) prior distributions on θ. When the likelihood of the data D, $L_D(\theta)$, defined by $P(D|\theta)$, whenever the lower marginal probability $\underline{E}(L_D) > 0$ and $f \in \text{Dom}(\underline{E}) = K \subseteq \mathcal{F}$, the posterior lower expectation,

$$\underline{E}(f|D) = \inf \ \left\{ \frac{E_P(f(\theta)L_D(\theta))}{E_P(L_D(\theta))} : \ E_P \in M \right\},$$

is jointly coherent with all likelihoods $L_D(\theta) = P(D|\theta)$ indexed over $\theta \in \Theta$ and $\underline{E}(\cdot)$.

□

Let us summarise the material up to this point. In Chapter 1, we discuss the need for models involving sets of distributions due to the nature of elicitation. At the beginning of this chapter, we discuss why avoiding losses and coherence is important when considering such models and that Theorems 2.2.1 and 2.2.2 justify the coherence of optimising expectations over a set of distributions. Thus the notion of coherence has been extended to include joint coherence between a set of conditional expectations and a set of unconditional expectations, with the GBR being one way of generating a conditional model that are jointly coherent with a given unconditional one. Finally, the GBR is applied to the context of prior/posterior inference. In all, we have the necessary tools to arrive at posterior imprecise models starting from elicitation of a set of prior distributions.

2.3 Imprecise Dirichlet Model (IDM)

The imprecise Dirichlet model (IDM) was first used by Walley [80] as an imprecise model for multinomial counts data. Mathematically, it is an application of Theorem 2.2.4 with M being a set of precise Dirichlet priors and with the single likelihood being the multinomial distribution. The imprecise Beta model in Examples 2.3.2 and 2.3.3 for binomial data was also introduced in Walley [81].

Fix the *concentration parameter* $\nu > 0$, $m \in \mathbb{N}$ and consider the set of candidate Dirichlet priors for a random vector $\boldsymbol{\theta}$,

$$M := \{\text{Dirichlet}(\nu\boldsymbol{\alpha}) : \ \boldsymbol{\alpha} \in \triangle^m\}.$$

For a suitably measurable random variable, $f(\boldsymbol{\theta})$, a dataset of multinomial counts, $\boldsymbol{n}$, Theorem 2.2.4 and the properties of the Dirichlet-Multinomial pairing yields the posterior lower expectation

$$\underline{E}_M(f|\boldsymbol{n}) = \inf \left\{E_{\text{Dirichlet}(\nu\boldsymbol{\alpha}+\boldsymbol{n})}(f) : \ \boldsymbol{\alpha} \in \triangle^m\right\},$$

Theorem 2.2.4 guarantees that this is coherent.

Example 2.3.1: (Posterior inference for first moment of a Bernoulli probability under a set of beta priors) Again, following the setting of Example 2.3.2, we can show, that when

P is Beta$(\nu\alpha, \nu(1-\alpha))$ for $\alpha \in (0,1)$, after observing a total of n observations and $n_1 \leq n$ number of 'successes', as the vector $\boldsymbol{n} = (n_1, n - n_1)$,

$$E_P(\theta|\nu, \alpha, n_1, n) = \frac{\nu\alpha + n_1}{\nu + n}.$$

The posterior lower expectation of θ is therefore

$$\underline{E}(\theta|\boldsymbol{n}) := \inf \{E_P(\theta|\nu, \alpha, n, n_1) : \ \alpha \in (0,1)\} = \inf_{\alpha \in (0,1)} \frac{\nu\alpha + n_1}{\nu + n} = \frac{n_1}{\nu + n}.$$

(Similarly, the upper expectation under this model is,

$$\overline{E}(\theta|\boldsymbol{n}) := \sup \{E_P(\theta|\nu, \alpha, n, n_1) : \ \alpha \in (0,1)\} = \sup_{\alpha \in (0,1)} \frac{\nu\alpha + n_1}{\nu + n} = \frac{\nu + n_1}{\nu + n}.$$

■

Example 2.3.2: (The imprecise Beta model, Walley [81]) For an observation space $\{s_1, s_2\}$, let $\theta \in [0,1]$ be a distribution on this space (such that $\Pr(\{s_1\}|\theta) = \theta$ and $\Pr(\{s_2\}|\theta) = 1 - \theta$). Walley [81] proposes to place a family of Beta priors on θ as follows. For a fixed *concentration parameter* $\nu > 0$, consider the set of Beta distributions,

$$\{B(\nu\alpha, \nu(1-\alpha)) : \ \alpha \in (0,1)\}.$$

Note that $\alpha \in (0,1)$ ensures the integrability of the distributions in this set. However, the case with $\alpha \in [0,1]$ is more appropriate in applications where one does not discount placing a prior probability of one on $\theta = 0$ and/or $\theta = 1$. We will explore this case more in Chapter 3.

■

Example 2.3.3: (IBM is vacuous for its positive integer non-central moments) Following the setting of Example 2.3.2, we can show, that when P is Beta$(\nu\alpha, \nu(1-\alpha))$ for $\alpha \in (0,1)$,

$$E_P(\theta|\nu, \alpha) = \alpha.$$

Therefore,

$$\underline{E}(\theta) := \inf \{E_P(\theta|\nu, \alpha) : \ \alpha \in (0,1)\} = 0 = \inf_{\theta \in [0,1]} \theta = 0.$$

(Similarly, the upper expectation under this model is 1.) The IBM is called a *near ignorant* model by Benavoli and Zaffalon [7]: for certain bounded random variables, their lower

and upper expectations are respectively their finite infimum and supremum. Because only the set of all distributions over a measurable space yield vacuous lower and upper expectations for all random variables in the measurable space [7], any restriction to this set of distributions should in principle mean that the lower and upper expectations for some random variable in the measurable space is non-vacuous. However, many common statistics are vacuous under these models. For example, for the Beta$(\nu\alpha, \nu(1-\alpha))$ distribution, expectations that are monotonic functions with respect to α, such as the positive moments,

$$E_P(\theta^k) = \prod_{i=1}^{k} \frac{\nu\alpha + i - 1}{\nu + i - 1},$$

for $k \geq 1$, are minimised at 0, by picking $\alpha = 0$ and maximised at 1, by picking $\alpha = 1$.

■

2.4 A commentary for statisticians

2.4.1 Properties of the IDM

Walley [80] notes that the inference of the IDM model carries forward the representation invariance from the precise Dirichlet-Multinomial model. That is, if $\theta \sim \text{Dir}(\alpha)$, any aggregation of a subset of indices of θ gives rise to a Dirichlet distribution with the corresponding α's summed, and this aggregation is done for every Dirichlet model in an IDM set. Walley argues for this principle in both [81] and [80]: however, some practitioners cite examples where this principle might not be required to hold in the discussions at the end of [80] (see the synthesis of responses to the IDM in Section 2.4.2). When the prior parameter set is taken to be $\{\nu\alpha : \alpha \in \triangle^K\}$ for a fixed concentration parameter $\nu > 0$, then inference is also invariant to permutations of categories (this is the symmetry principle [81]). Bernard [13] notes many proper priors subsumed by the precise Dirichlet model (for example [13] $\nu = K$, $\nu = 1$ and $\nu = K/2$, respectively the Laplace uniform, Perks [64] and Jeffreys [50] priors) have posteriors whose probability of an event depends on on how the event is represented with respect to the set of multinomial categories. Haldane's prior [44], with $\nu \to 0$, satisfies these principles, but does not avoid sure loss. The IDM is therefore motivated as a model which satisfies all of these criteria.

2.4.2 A synthesis of responses to IDM from statistical community

We synthesise some comments from the statistical community regarding the principles upon which the IDM is built, as well as some of its properties. We review some relevant comments in the discussions at the end of Walley [80]. Except for explicit citations, all other author references in this section refer only to those appearing in Walley [80].

Throughout Walley [80], the IDM is applied to two examples. The first is of a conceptual bag of marbles representing inference of a multinomial sampling distribution, with the goal of inferring the prior probability of drawing a red marble (R) without knowing anything about the composition inside the bag. This example is used to demonstrate that no single (precise) Bayesian model can produce a coherent prior inference (without additional assumptions.) The IDM is proposed as an alternative that generates reasonable prior inference (namely, $\underline{P}_{\text{IDM}}[R] = 0$ and $\overline{P}_{\text{IDM}}[R] = 1$), while producing non-trivial, reasonable and coherent posterior inference as data is accumulated. The IDM is then applied in a hypothesis test setting to the extracorporeal membrane oxygenation (ECMO) dataset to determine whether or not the ECMO treatment is better than the conventional treatment.

ν as a tuning parameter: discussants observed that there is no principled way of choosing the value ν, and therefore, as Good notes, makes the IDM more of a subjective than objective (extension of a) Bayesian model. Furthermore, Levi notes that "choosing a value for $[\nu]$ is one of several ways of exercising boldness in forming prior opinions", and that "for $[\nu]= 1$ and $[\nu]= 2$, the inquirer is far from full probabilistic ignorance." (This is interesting as Walley proposes that these are suitable heuristics, but is sceptical that they cannot be justified in principle.) Hutton questions whether or not such an arbitrary choice representing prior subjectivity should be represented merely by a one-dimensional object. Coolen gives an example of interest to us. If one was at a mall, and one encounters a person named John, then, according to the IDM, the lower probability of randomly encountering another John is $\underline{P}_{\text{IDM}} = 1/(1+\nu)$, that, for $\nu = 1, 2$ the probability of this happening is at least $1/2$ and $1/3$, respectively. Coolen deems this to be too high for a single observation.

Representation invariance is not always valid: discussants were concerned that representation invariance (that the inference of the IDM does not depend on the representation of the sample space, which was a criticism levelled at many precise Bayesian models by Walley in the body of the paper), should not always hold. For example, Dawid notes that coarsening of the multinomial bins typically discards information, and should therefore result in a different inference, and, if structural knowledge about the probabilities such as the possibility that they vary continuously over a real variate that has been discretised

by binning (such as bins of heights of people), then a smoothed distribution should be expected to produce more realistic and reasonable inference, which is not possible if the model is representation invariant. Walley concedes to the idea that not all models should be representation invariant, but nevertheless notes that representation invariance is desirable when prior knowledge such as the ones discussed by Dawid are not available.

Over complexity of the IDM (and lower and upper probabilities in general): some discussants have claimed that the IDM may be replaced by simpler, more familiar precise Bayesian paradigm. For example, O'Hagan and Lindley state that it is possible to elicit a proper prior under some reasonable assumptions about the bag of marbles example. Walley's rejoinder to this is that it is unnecessary (and unjustified given no prior knowledge) for coherent prior inference to assign a particular single set of probabilities which are conditional on the knowledge of a fixed sample space of colours to begin. (Note that this was made in reference to the use of imprecise probabilities in general.) Walker cited that the generalised Polya urn model provides the same lower and upper posterior expectations of a cell probability as the IDM (although it is unclear if this holds for other functions of the cell probabilities.) Finally, Spiegelhalter and Best argue that a (local) sensitivity analysis is more appropriate in the ECMO example. Walley's rejoinder to these points is that (local) sensitivity analysis only makes sense when "some meaning can be given to the 'correct' prior distribution and there is uncertainty about which distribution is correct". Finally, Walley notes that, for the bag of marbles example, it would be even more complex to elicit a single distribution that can take into account of any and all possible compositions in the bag.

Conservatism of imprecise probabilities and the need for precise decisions: some discussants point out that imprecise models sometimes provide inference that does not necessarily provide a definite conclusion. For example. Lindley questions how a medical decision should proceed when the (imprecise) model does not give a definite answer to distinguish between, say, two treatments. Walley replies that, in this case, the model indicates that both treatments are equally preferable, and therefore random allocation can be used to assign treatments.

2.4.3 A brief review of statistics in imprecise probabilities

We choose to work with Walley's framework primarily as a gateway in our exploration of imprecise probabilities. Because Walley's work generalises a vein of Bayesian statistics in an apparent manner (i.e. through the lower envelope version of the generalised Bayes' rule, Proposition 2.2.1), our choice is meant merely to leverage this familiar grounding to

explore imprecision from the author's "precise statistical" perspective. By no means is our perspective closed off towards other approaches towards imprecision.

Research progress has been done by the imprecise probability community towards applying imprecise probability towards statistical concerns. A select list of such work include the following.

Weichselberger [83] introduced a generalisation of Kolmogorov's measure theoretic probabilities, known as *interval probability* (or *F-probability*). Augustin and Coolen [6] combined F-probabilities with Hill's nonparametric inference (NPI) ([46] and [45]) to form an imprecise probabilities model with frequentist properties. (In particular, Coolen and Augustin [23] compare this model with the IDM.) Recently, much work has been done on the application of this imprecise version of NPI towards statistical hypothesis testing and prediction problems in various publications of T. Maturi-Coolen and her collaborators (such as [25], [58], [20], [4] and [24]). Benavoli, Magnili, Ruggeri and Zaffalon [8] introduced a generalisation of the IDM to the *imprecise Dirichlet process* and applied it to a hypothesis test regarding a cell probability of the multinomial likelihood. Perolat, Couso, Loquin and Strauss [65] generalised the Mann-Whitney U test for random set data (so-called *imprecise(ly observed) data* by some researchers). Couso and Dubois [27] generalised the maximum likelihood procedure for imprecise data. Cattaneo and Wiencierz [19] introduce a likelihood based approach to regression with imprecise data. Troffaes [76] explored several criteria towards loss minimisation during decision-making when using imprecise probabilities. Finally, Dempster's rule of combination in the Dempster-Shafer theory (originally proposed by Dempster [34] and later further developed in Shafer [73]) has been a long standing methodology in the imprecise probabilities as well as the sensor fusion literature.

The contributions in this thesis seek to be complementary to the above work. We place further emphasis and attention towards the exploration of the following items.

- The incorporation of stochastic variation (such as credible and confidence interval statistics from a precise distribution) with variations due to imprecision (such as the posterior imprecise expectations produced by Theorem 2.2.4).
- The interplay between the occurrence of sparse data counts in multinomial observations and a principled approach to sensitivity analysis (for example, under Walley's coherence).

- The statistical interpretability of posterior inference in the above contexts, especially in light of Walley's coherence.
- The computational tractability of the IDM when applied to statistics beyond the multinomial cell probability (particularly in this thesis, to log-odds statistics).

The first point is crucial for a complete generalisation and subsumption of the inferential methodology with a single precise distribution into the imprecise methodology. The third and fourth points highlight important obstacles against the mainstream statistical adoption of imprecise probabilistic methodology. The lack of an established statistical interpretation sours the prospect of having to overcome the optimisation problem during the computation of imprecise expectations. A statistician committee member of this thesis summarised this concisely. The computation for Bayesian posterior expectations is difficult enough without having to optimise it as a function of the prior. Prospective statisticians considering using imprecise probabilities (in the form of Theorem 2.2.4) need to be given a very compelling reason to undertake this expensive operation. This thesis attempts to explore whether or not such reasons exist in application.

2.4.4 Comments on imprecise models

We end this introductory chapter with a commentary on some key points, that we ourselves found helpful during the course of writing this document, for the reader to keep in mind while reading the rest of the document.

Imprecision versus uncertainty: It is important to distinguish between the concepts of uncertainty and imprecision. Although they both induce variation in inference that statisticians will take into account, we should distinguish the sources. Uncertainty comes from the randomness that is modelled by a single precise distribution: variation of this kind is usually interpreted in ways such as the variation represented by frequentist confidence and Bayesian credible intervals. On the other hand, this is distinct from imprecision where the variation is due not to the randomness of the system, but our *inability to specify the randomness of the system.*

Example 2.4.1: To illustrate this point, for a random variable $\theta \in [0,1]$, consider a set of Dirac delta distributions over $[0,1]$ that it can take as law;

$$M = \{\delta_a(\cdot) : \ a \in [0,1]\}.$$

Then, it is clear that the variance of each distribution is 0, while, for example, the lower and upper expectations are 0 and 1. We see that the former is 'within' variation of due to the uncertainty represented by each distribution, while the latter is variation and property 'amongst' the distributions of a set.

■

Imprecision versus sensitivity analysis: Recall that the distinction between the range in (global) sensitivity analysis reviewed in Chapter 1 and imprecision defined here as part of the imprecise methodology. While they both compute the difference between the maximum and minimum of expectations, they are interpreted very differently as sensitivity analysis is not *defined* to be coherent, while imprecise probabilities and expectations have such a notion. This means that the latter can be used as part of the inference in the same way coherent probabilities can be without inducing logical inconsistencies such as Dutch book assessments. This is in contrast to a sensitivity analysis being only treated as a post-hoc methodology separate from the inferential engine of coherent probabilities. Again, the point of this thesis is to study aspects of the consequence of using sets of distributions as a part of the inference through the imprecise methodology.

Boundedness of random variables and coherence: Because the notion of coherence for imprecise expectations follows Walley [81], it is motivated by betting procedures that are troublesome to interpret when the random variable is unbounded. For an in-depth treatment of extensions to extended-real-valued (and, therefore, unbounded) random variables in the theory of imprecise probabilities, see Troffaes and de Cooman [78]. Parts of this thesis will involve only bounded random variables. In others, such a Chapter 3, part of our contributions is to justify certain limiting procedures as a sound methodology and explore their consequences.

Chapter 3

Log-odds inference under IDM and sparse observations

In this chapter, we apply the *imprecise Dirichlet model* (IDM, Walley [81] [80]) to the posterior inference of log-odds statistics of a multinomial system. We motivate the need for imprecision by first showing that posterior inference can be sensitive to the choice of prior when the likelihood does not dominate the prior in posterior expressions. This can occur when the multinomial counts contain zero counts in certain categories, and we focus on inference when this kind of sparsity of observations occur. The set of distributions of the IDM captures this notion of prior sensitivity. Bickis [14] notes that the geometrical properties in the natural parameter space of exponential families can be used to easily construct imprecise posterior inference. However, under our sparse cases, we will see that the boundary of this space comes into effect and requires careful consideration.

A significant part of our work in this chapter deals with a modification of the IDM under Walley's theory in order to apply it to the log-odds statistic. This is in order to overcome the fact that Walley's imprecise probabilities are only defined for *bounded* random variable, whereas the log-odds statistic is *unbounded*. Walley's [81] notion of coherence introduced in Chapter 2 breaks down in light of this: we will discuss the relation between our modifications and Walley's notion of coherence.

3.1 Sensitivity of posterior inference to prior choice

Under the multinomial setting, the posterior inference is sensitive to prior choice when the likelihood is almost or exactly flat in certain directions that extend to the boundary of the

parameter space. Particularly, distributional approximations such as Gaussian approximations become increasingly poor close to the boundary. In these cases, sensitivity analysis is necessary to analyse the effects of the prior on inference where the likelihood is relatively uninformative.

Below we highlight these points in the Dirichlet-Multinomial setting.

Example 3.1.1: For a Dirichlet prior with hyperparameter $\nu(\alpha_1, \alpha_2, 1-\alpha_1-\alpha_2)$, $\nu > 0$, after observing the i.i.d.-trinomial data $\boldsymbol{n} = (n_1, n_2, n_3)$, the posterior has the kernel,

$$(\theta_1, \theta_2) \longmapsto \theta_1^{n_1+\nu\alpha_1-1}\theta_2^{n_2+\nu\alpha_2-1}(1-\theta_1-\theta_2)^{n-n_1-n_2+\nu(1-\alpha_1-\alpha_2)-1}.$$

where $(\theta_1, \theta_2, 1-\theta_1-\theta_2)$ is a trinomial probability.

Let us consider using a single fixed Dirichlet prior with concentration $\nu = 2.0$ and $(\alpha_1, \alpha_2, 1-\alpha_1-\alpha_2) = (0.01, 0.98, 0.01)$, and consider the cases where we observe two different datasets $(n_1, n_2, n_3) = (10, 1, 10)$ and $(10, 0, 10)$. Notice that the prior direction puts a lot of weight to the second category, whereas the dataset suggests otherwise in both cases.

Consider Figure 3.1, the case with the dataset $(10, 0, 10)$ which has 0 samples in the second cell. A prior Dirichlet concentration parameter of 2.0, when interpreted as the effective prior sample size, is weak relative to a dataset with 20 observations. Yet it is strong enough to force the posterior contours to be unimodal, even if it is not well approximated by a Gaussian. This is in contrast with the likelihood (lower panel of figure) which flats out to the bottom direction. This demonstrates the sensitivity of posterior inference to prior choice in this setting. We note that this is despite the prior being relatively weak compared to the total number of observations.

We see that sparsity (i.e. the observed 0 in the second cell of $(n_1, n_2, n_3) = (10, 0, 10)$) is a significant contributor to this sensitivity. In Figure 3.2, we see the significant effect of just adding 1 observation to cell 2: the likelihood contours are mostly pulled back to the centre, just like those of the posterior. So, sparsity in a cell category represents a major contributor to the sensitivity to the prior in Dirichlet-Multinomial systems.

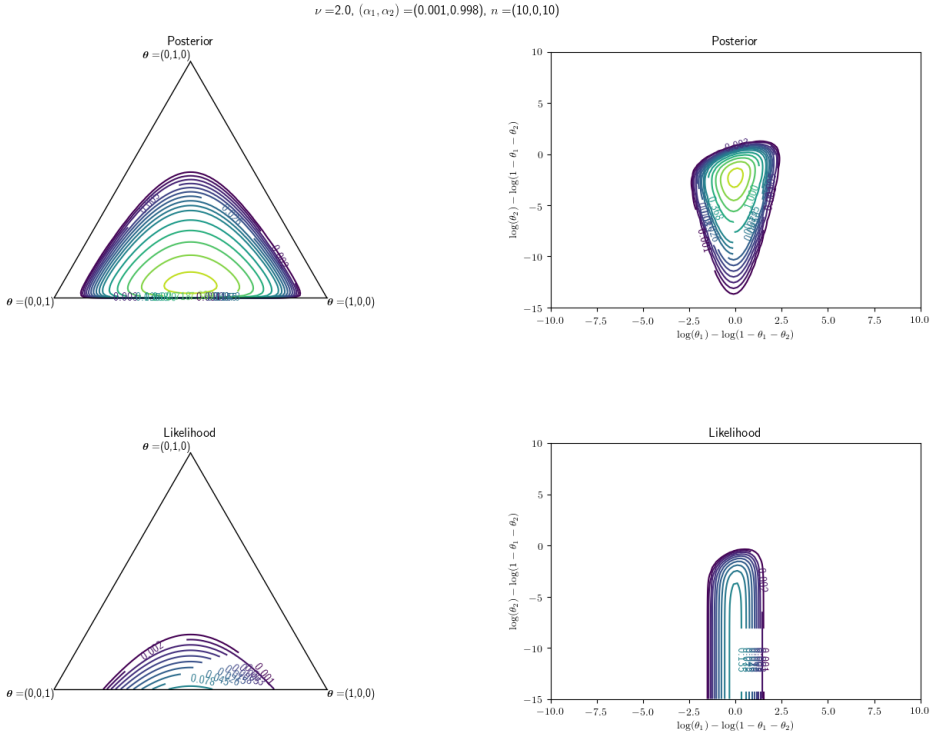

Figure 3.1: Contour plots of the posterior density $p((\theta_1, \theta_2, 1 - \theta_1 - \theta_2)|\nu(\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2) + (n_1, n_2, n_3))$ (Top) and the trinomial likelihood $L((n_1, n_2, n_3)|(\theta_1, \theta_2, 1 - \theta_1 - \theta_2))$ (Bottom). Plots are in barycentric coordinates relative to the space of trinomial distributions $\text{Conv}(\{(0, 0, 1), (0, 1, 0), (1, 0, 0)\})$ (left) and in log-odds space of all trinomial distributions (right). Posterior parameters are $\nu = 2.0$, $(\alpha_1, \alpha_2) = (0.001, 0.998)$ and $(n_1, n_2, n_3) = (10, 0, 10)$.

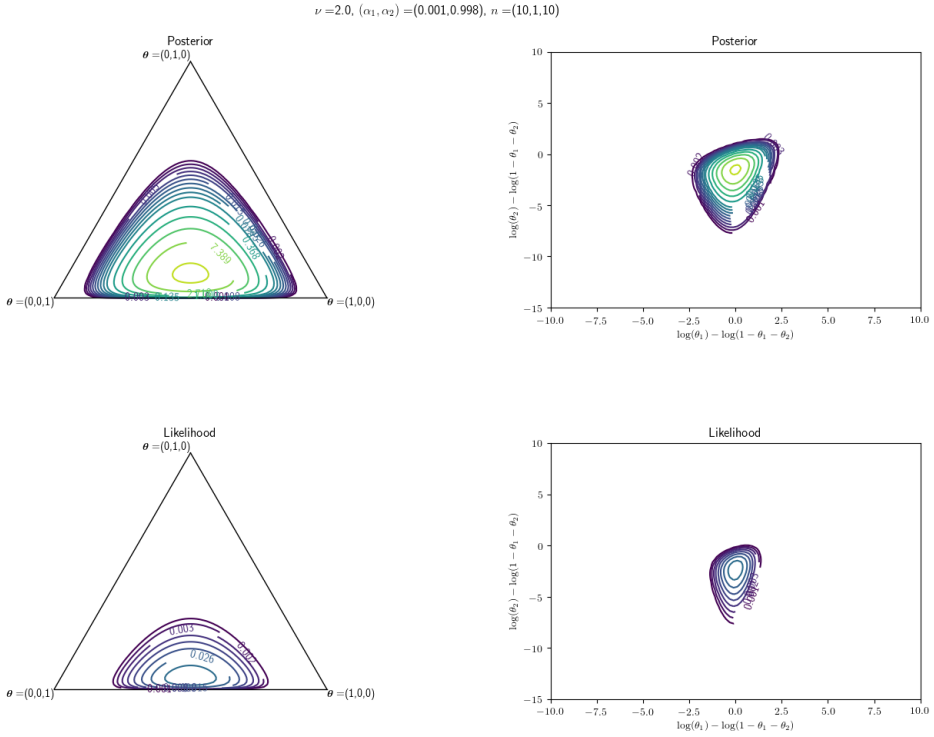

Figure 3.2: Contour plots of the posterior density $p((\theta_1, \theta_2, 1 - \theta_1 - \theta_2)|\nu(\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2) + (n_1, n_2, n_3))$ (Top) and the trinomial likelihood $L((n_1, n_2, n_3)|(\theta_1, \theta_2, 1 - \theta_1 - \theta_2))$ (Bottom). Plots are in barycentric coordinates relative to the space of trinomial distributions $\text{Conv}(\{(0, 0, 1), (0, 1, 0), (1, 0, 0)\})$ (left) and in log-odds space of all trinomial distributions (right). Posterior parameters are $\nu = 2.0$, $(\alpha_1, \alpha_2) = (0.001, 0.998)$ and $(n_1, n_2, n_3) = (10, 1, 10)$.

■

This example illustrates that in the sparse data case, when likelihood information does not uniformly dominate prior information, we expect strong sensitivity to prior specification. In particular, as shown in Chapter 1, we recall that elicitation often produces imprecise prior specification. So we expect imprecise posterior specification here. This Chapter explores if the tools of imprecise inference, described in Chapter 2, can be helpful to the statistician in this situation.

In our exploration we take inference about log-odds, in the IDM framework, as a practical and statistically important test case for the applications of imprecise methods. Because the log-odds is unbounded, we see that in Section 3.3.1 the boundedness condition on random variables used in Chapter 2 immediately creates a problem for application.

3.2 Literature: Affine geometry of IDM posterior updating under sparse observations

We apply the work of Bickis [14] to establish that updating the posterior of the IDM amounts to the translation of a simplex in the prior Dirichlet natural parameter space by the observation vector which is a sufficient statistic of the likelihood. We see that in the sparse case the translation does not move the posterior support entirely into the relative interior and so the boundary still plays a role in understanding the sensitivity problem.

3.2.1 Affine geometry of exponential family and Dirichlet-multinomial updating

Following Bickis [14], we review a familiar result of conjugate pairs of prior and observation models.

Definition 3.2.1: A random variable X follows a distribution in the *exponential family spanned by the sufficient statistic* $\boldsymbol{v}(x) = (v_1(x), \ldots, v_k(x))$ *(with respect to a base measure* λ*)* if there exists a vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)$ such that,

$$P(X \in A) = \int_A \exp(\boldsymbol{\xi}^T \boldsymbol{v}(x) - \phi(\boldsymbol{\xi}))\lambda(dx),$$

where,

$$\phi(\boldsymbol{\xi}) = \log\left(\int \exp(\boldsymbol{\xi}^T \boldsymbol{v}(x))\lambda(dx)\right).$$

The family is said to be *finite dimensional* when $k < \infty$.

□

Definition 3.2.2: The *natural parameter space* of an exponential family spanned by $\boldsymbol{v}$ is the set,

$$\Xi_{\text{Nat}} = \{\boldsymbol{\xi} : \ \phi(\boldsymbol{\xi}) < \infty\}.$$

□

Now, consider that a random variable X which follows a member of a finite dimensional exponential family spanned by $\boldsymbol{v} = (v_1(x), \ldots, v_k(x))$ with natural parameters $\boldsymbol{\xi}$. Place a prior on $\boldsymbol{\xi}$ with log-density of the exponential form,

$$\boldsymbol{\xi} \longmapsto \boldsymbol{\eta}^T \boldsymbol{v}^*(\boldsymbol{\xi}) - \psi(\boldsymbol{\eta}), \tag{3.1}$$

spanned by,

$$\boldsymbol{v}^*(\boldsymbol{\xi}) = (-\phi(\boldsymbol{\xi}), \xi_1, \ldots, \xi_k),$$

with natural parameters,

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_k),$$

For a fixed data point, x, of X,

$$\boldsymbol{\xi}^T \boldsymbol{v}(x) - \phi(\boldsymbol{\xi}) = \boldsymbol{v}^*(\boldsymbol{\xi})^T \boldsymbol{\eta}_x,$$

where,

$$\boldsymbol{\eta}_x = (1, v_1(x), \ldots, v_k(x)).$$

Now, consider the posterior measure after observing x with the aforementioned prior natural parameters $\boldsymbol{\eta}$. For every suitably measurable set B,

$$\begin{aligned}\Pi(\boldsymbol{\xi} \in B|x) &= k(x) \int_B \exp(\boldsymbol{\xi}^T \boldsymbol{v}(x) - \phi(\boldsymbol{\xi})) \exp(\boldsymbol{\eta}^T \boldsymbol{v}^*(\boldsymbol{\xi}) - \psi(\boldsymbol{\eta})) d\boldsymbol{\xi} \\ &= \int_B \exp((\boldsymbol{\eta}_x + \boldsymbol{\eta})^T \boldsymbol{v}^*(\boldsymbol{\xi}) - \psi(\boldsymbol{\eta}) + \log k(x) d\boldsymbol{\xi}.\end{aligned}$$

So, the posterior and prior are conjugate and the updating result in a translation in the natural parameters. We record this in the theorem below.

Theorem 3.2.1: (Bickis [14]) For a likelihood of i.i.d. data $\{x_1, \ldots, x_n\}$ induced by an exponential family naturally parametrised by $\boldsymbol{\xi} \in \Xi_{\text{Nat}} \subseteq \mathbb{R}^k$, $k < \infty$, spanned by the

statistics $\boldsymbol{v}(x) = (v_1(x), \ldots, v_k(x))$ and a prior distribution of the form in (3.1), the posterior distribution is again in the exponential family of the prior, with natural parameters,

$$\left(\eta_0 + n, \eta_1 + \sum_{j=1}^{n} v_1(x_j), \ldots, \eta_k + \sum_{j=1}^{n} v_k(x_j)\right).$$

□

Example 3.2.1: (Dirichlet-categorical posterior) A Dirichlet prior on $\boldsymbol{\theta}$ with parameters ν and $\boldsymbol{\alpha}$ over m categories can be identified with the following terms of the exponential form: for $i = 1, \ldots, m-1$,

$$v_0^*(\boldsymbol{\theta}) = \log(1 - \theta_1 - \ldots - \theta_{m-1}),$$

$$v_i^*(\boldsymbol{\theta}) = \log \theta_i / (1 - \theta_1 - \ldots - \theta_{m-1}),$$

$$\eta_0 = \nu, \quad \eta_i = \nu\alpha_i, \quad \psi(\boldsymbol{\eta}) = \log B(\nu\boldsymbol{\alpha}) \ .$$

($B(\cdot)$ is the multivariate Beta function.) A categorical likelihood can also be identified with the following exponential form specification:

$$v_i(x) = I(x = i), \quad \xi_i(\boldsymbol{\theta}) = \log \tfrac{\theta_i}{1-\theta_1-\ldots-\theta_{m-1}}, \quad \phi(\boldsymbol{\theta}) = \log(1 - \theta_1 - \ldots - \theta_{m-1}) \ .$$

By Theorem 3.2.1, the posterior after $n < \infty$ i.i.d. observations is also Dirichlet, with the natural parameter,

$$\boldsymbol{\eta} + \boldsymbol{\eta}_{\boldsymbol{x}} = (\nu + n, \nu\alpha_1 + n_1, \ldots, \nu\alpha_{m-1} + n_{m-1}),$$

which is the well-known Dirichlet-multinomial updating property.

■

3.2.2 Affine geometry of the IDM

The update rule in Theorem 3.2.1 makes posterior lower expectations under the envelope version of the generalised Bayes' rule involving exponential families easy to characterise. Write

$$\overline{\blacktriangle^{m-1}} := \left\{(\alpha_1, \ldots, \alpha_{m-1}) : \ \alpha_i \geq 0, \ \sum_{i=1}^{m-1} \alpha_i \leq 1\right\},$$

to denote the closure of the interior of the simplex in $\mathbb{R}^m$.

Example 3.2.2: For a fixed concentration $\nu > 0$ and a count vector of n i.i.d. observations from m categories parametrised as $\boldsymbol{n} = (n, n_1, \ldots, n_{m-1})$, the posterior set of distributions of the IDM is,

$$M_{|\boldsymbol{n}} = \left\{ \text{Dirichlet}(\nu + n, \nu\alpha_1 + n_1, \ldots, \nu\alpha_{m-1} + n_{m-1}) : \ (\alpha_1, \ldots, \alpha_{m-1}) \in \overline{\blacktriangle^{m-1}} \right\}.$$

■

In particular, Theorem 3.2.1 implies that posterior update amounts to an affine shift of the prior coordinates by the sufficient statistics of the observation model. In other words, it can be seen that the natural parameters of the IDM posterior set of distributions is simply a translation of the whole prior set,

$$\{(\nu, \nu\alpha_1, \ldots, \nu\alpha_{m-1}) : \ (\alpha_1, \ldots, \alpha_{m-1}) \in \overline{\blacktriangle^{m-1}}\} \oplus (n, n_1, \ldots, n_{m-1}). \tag{3.2}$$

This is such that the *geometry* of posterior update of the IDM is an affine translation of a *fixed* closed simplex in the extended natural parameter space.

Notice that, at every fixed number of total observations n, the set of posterior natural parameters, (3.2), is always contained in the set of posterior parameters over all possible ways to observe the counts $(n, n_1, \ldots, n_{m-1})$. This is simply the set of all Dirichlet parameters that sum to $\nu + n$: that is, it is the simplex

$$\{(\gamma_1, \ldots, \gamma_m) : \ \gamma_i > 0, \ \gamma_1 + \ldots + \gamma_m = \nu + n\}. \tag{3.3}$$

This particular containment implies the following boundary effects when optimising over $M_{|\boldsymbol{n}}$ whose parameters are taken from (3.2).

Theorem 3.2.2: When $n_i > 0$ for all $i = 1, \ldots, m$, the translated set (3.2) is always in the interior of (3.3), whereas when $n_i = 0$ for some $i = 1, \ldots m$, the posterior translated set will intersect and travel along the face of (3.3) where $\alpha_i = 0$.

□

We will explore the consequences of inference with a set of distributions that intersects the topological boundaries of the natural parameter space of the Dirichlet family.

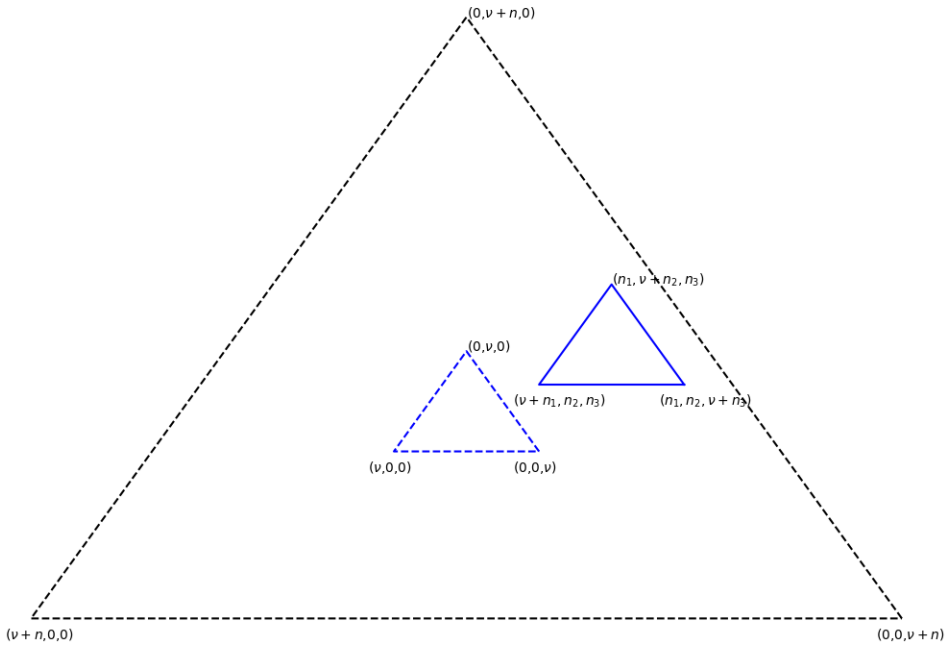

Figure 3.3: A geometrical view of IDM update by translation. The larger simplex (dashed black) represents the set (3.3) of possible Dirichlet posteriors after observing n observations with ν fixed apriori. The simplex of size ν (dashed blue) represents the natural parameters of the prior IDM set of distributions and the translation of this simplex by $\boldsymbol{n} = (n_1, n_2, n_3)$ (solid blue) represents the natural parameters (3.2) of the posterior IDM set of distributions.

3.3 Inference for log-odds under IDM

We use categorical data analysis as a backdrop to investigate statistical inference using imprecise probabilities. We focus on the imprecise IDM version of the following log-odds inference. In the precise case with a single prior, under the Dirichlet($\nu\boldsymbol{\alpha}+\boldsymbol{n}$) posterior distribution over m categories, for two collections of subsets of cell categories, $\mathcal{A} = \{A_1, \ldots, A_r\}$ with each $A_i \subseteq \{1, \ldots, m\}$ and $\mathcal{B} = \{B_1, \ldots, B_s\}$ with each $B_j \subseteq \{1, \ldots, m\}$, the expectation of the general log-odds of the following form is a linear combination of digamma functions,

$$E\left(\log \frac{\theta_{A_1} \ldots \theta_{A_r}}{\theta_{B_1} \ldots \theta_{B_s}} \,\middle|\, \nu\boldsymbol{\alpha} + \boldsymbol{n}\right) = \sum_{i=1}^{r} \psi(\nu\alpha_{A_i} + n_{A_i}) - \sum_{j=1}^{s} \psi(\nu\alpha_{B_j} + n_{B_j}) - (|\mathcal{A}| - |\mathcal{B}|)\psi(\nu + n). \tag{3.4}$$

where for a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, and a subset of indices $C \subseteq \{1, \ldots, m\}$,

$$\theta_C := \sum_{i \in C} \theta_i. \tag{3.5}$$

(This follows from the fact that the log-probabilities $\log \theta_i$ are sufficient statistics of the Dirichlet distribution and the digamma functions result from differentiating the Beta function of the normalising constant.) The IDM version of this involves computing the posterior lower expectation,

$$\underline{E}_{\text{IDM}}\left(\log \frac{\theta_{A_1} \ldots \theta_{A_r}}{\theta_{B_1} \ldots \theta_{B_s}} \,\middle|\, \nu, \boldsymbol{n}\right) = \inf\left\{E\left(\log \frac{\theta_{A_1} \ldots \theta_{A_r}}{\theta_{B_1} \ldots \theta_{B_s}} \,\middle|\, \nu\boldsymbol{\alpha} + \boldsymbol{n}\right) : \boldsymbol{\alpha} \in \triangle^m\right\}. \tag{3.6}$$

From an optimisation perspective, we note that, when counts of some observation cells, n_i's, are zero, the objective function above contains terms of the form $\psi(\nu a)$ which tends to $-\infty$ as the optimisation variable a tends to 0. We will examine this in more details in Chapter 5.

3.3.1 Unboundedness of the log-odds and the theory of coherence

We are interested in the cases where the log-odds statistic becomes unbounded when $\boldsymbol{\theta}$ contain zeroes as prior sensitivity may be extreme. However, this case presents a problem for directly applying the mathematical framework in Chapter 2 as Walley's coherence of lower expectations [81] are defined only on sets of bounded random variables. There has

been relatively little work on extensions to unbounded random variables: we cite Troffaes and de Cooman [77], Troffaes [75], their culmination in the later book by Troffaes and de Cooman [78] as well as the investigations by Crisma, Gigante, and Millossovich [31] and Schervish, Seidenfeld and Kadane [71].

We take an approach that is more in line with sensitivity analyses. We restrict our scope of analysis to the IDM and justify our use of it with the log-odds with limiting arguments. Specifically, we consider a truncation of the log-odds that is bounded and finite.

Definition 3.3.1: For a coherent lower expectation, $\underline{E}$, defined over the set of bounded random variable $\mathcal{L}(\Omega)$s, write $\underline{E}^{(\text{ext})}$ as its extension to the linear space,

$$\text{span}(\mathcal{L}(\Omega) \cup \{g\}).$$

□

When the multinomial is non-sparse in the sense that each category has at least one observation, we show that the approximation error between the posterior expectations of a truncated log-odds and the original one converges to zero in the L^1 sense uniformly over the posterior natural parameter space that is the optimisation domain of the IDM.

Theorem 3.3.1: (Theorem B.2.1) Consider the general log-odds statistic,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

and the following truncation of it,

$$T_c(\boldsymbol{\theta}) := g(\boldsymbol{\theta}) I(|g(\boldsymbol{\theta})| \leq c) + cI(|g(\boldsymbol{\theta})| > c).$$

Under $\boldsymbol{\theta} \sim \text{Dirichlet}(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ with sets $A_1, \ldots, A_r, B_1, \ldots, B_s$, such that $n_{A_i} > 0, n_{B_j} > 0$ for all A_i, B_j,

$$\sup_{\boldsymbol{\alpha} \in \overline{\triangle^{|\Omega|}}} E(|T_c - g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

□

The coherence of Troffaes and de Cooman [78] extends that of Walley [81] (introduced in Chapter 2) to unbounded ones such as the general log-odds by using an approximation scheme. It can be shown that our construction is also coherent under their extended coherence notions.

Theorem 3.3.2: (Theorem B.2.2) Under the conditions of Theorem 3.3.1, the extended IDM whose value at the unbounded log-odds g is given by,

$$\underline{E}^{(\text{ext})}_{\text{IDM}}(g(\boldsymbol{\theta})|\nu,\boldsymbol{n}) = \lim_{c\to\infty} \underline{E}_{\text{IDM}}(T_c(\boldsymbol{\theta})|\nu\boldsymbol{\alpha}+\boldsymbol{n}),$$

is coherent under the Proposition B.2.1 of Troffaes and de Cooman [78].

□

Following their extension allows us to make the following interpretation of performing the optimisation and treating its result as an extension of the IDM lower expectation to include g in its domain.

Interpretation 3.3.1: For any general log-odds,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

whenever $\boldsymbol{n}$ contains at least one count in each category, we interpret, $\underline{E}^{(\text{ext})}_{\text{IDM}}(g(\boldsymbol{\theta})|\nu,\boldsymbol{n})$ as a double limit of the sequence of truncated Dirichlet expectations,

$$c,\boldsymbol{\alpha} \longmapsto E(T_c|\nu\boldsymbol{\alpha}+\boldsymbol{n}) = E\left(g(\boldsymbol{\theta})I(|g(\boldsymbol{\theta})|\leq c) + cI(|g(\boldsymbol{\theta})>c|)\,\middle|\,\nu\boldsymbol{\alpha}+\boldsymbol{n}\right),$$

over the optimisation path of $\boldsymbol{\alpha}$ contained in $\overline{\triangle^{\Omega}}$ and the limit of the truncation approximation via $c\to\infty$.

As described in Chapter 2, we can interpret the imprecise theory as either a sensitivity analysis of elicited priors or in terms of coherence of direct assessments about random variables. However, care must be taken when applying the interpretation of Section 2.1 motivating the coherence definition 2.1.8 to $\underline{E}^{(\text{ext})}_{\text{IDM}}(g|\nu,\boldsymbol{n}) = -\infty$ (and, by conjugacy, $\overline{E}^{(\text{ext})}_{\text{IDM}}(g|\nu,\boldsymbol{n}) = \infty$). See 13.11 in Troffaes and de Cooman [78].

3.3.2 The divergence of coherence from sensitivity analysis under sparse observations

Despite the material in the previous section, the behavioural and mathematical extension of Walley's coherence [81] for the IDM encounters further obstacles when it is used to perform posterior inference for the log odds-type statistics. In particular, we now demonstrate that the error of the previous approximation does not converge to zero if certain cells have zero samples.

Example 3.3.1: Consider $\boldsymbol{\theta} \sim \text{Dirichlet}(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2) \in \triangle^{|\Omega|}$ and,

$$g(\boldsymbol{\theta}) = \log \theta_1/\theta_2,$$

for $\boldsymbol{\theta}$ a vector of trinomial cell probabilities.

Consider the case when $\boldsymbol{n} = (0, 0, n_3)$ such that $n_3 > 0$:

$$\begin{aligned}
&\lim_{\alpha_1 \to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) \\
&\geq \left(\lim_{\alpha_1 \to 0} E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(|g|I(|g| \leq c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(cI(|g| > c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \\
&= \lim_{\alpha_1 \to 0} \left(E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(|g|I(|g| \leq c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) - cP(|g| > c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \\
&> \lim_{\alpha_1 \to 0} E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - 2c \\
&\geq \lim_{\alpha_1 \to 0} |E(g|\nu\boldsymbol{\alpha} + \boldsymbol{n})| - 2c \\
&= \lim_{\alpha_1 \to 0} |\psi(\nu\alpha_1) - \psi(\nu\alpha_2)| - 2c \qquad (\psi \text{ is the digamma function}) \\
&= \infty,
\end{aligned}$$

Similarly,

$$\lim_{\alpha_2 \to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) \geq \infty.$$

This leads to,

$$\begin{aligned}
&\lim_{c \to \infty} \lim_{\alpha_1 \to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = \infty \\
&\lim_{c \to \infty} \lim_{\alpha_2 \to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = \infty
\end{aligned} \quad . \tag{3.7}$$

That is, the error of the truncation approximation does not converge to zero in this case.

■

Example 3.3.1 demonstrates that, in the case of interest where certain cells have zero observations, there are issues with Walley's coherence [81] on bounded random variables and with its extension by Troffaes and de Cooman [78] to unbounded random variables via approximating random variables for posterior inference of log-odds.

Nevertheless, if we put coherence aside, the optimisation,

$$\inf_{\boldsymbol{\alpha}\in\triangle^{|\Omega|}} E_{\text{Dir}}(g|\nu\boldsymbol{\alpha}+\boldsymbol{n}),$$

itself can still yield well-defined solutions. This is because, for the general log-odds,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

its expectation under any posterior Dirichlet distribution in the set is a linear combination of digamma functions,

$$E_{\text{Dir}}(g|\nu\boldsymbol{\alpha}+\boldsymbol{n}) = \sum_{i=1}^{r} \psi(\nu\alpha_{A_i}+n_{A_i}) - \sum_{j=1}^{s} \psi(\nu\alpha_{B_j}+n_{B_j}) - (r-s)\psi(\nu+n),$$

and is therefore a smooth function in the interior of the $\boldsymbol{\alpha}$ simplex. (Recall that $a_C = \sum_{i\in C} a_i$ for a vector $\boldsymbol{a}$ and a set of indices C, and n is the total number of observations.) In the following sections, Theorem 3.4.1 show that this expectation may optimise to $\pm\infty$ for when either all the sets of the numerator or all the sets of the denominator of $g(\boldsymbol{\theta})$ have strictly positive cell counts. This yields an interpretable inference in the vein of a global sensitivity analysis methodology.

Interpretation 3.3.2: When there exists no $A_i \in \mathcal{A}$ and $B_j \in \mathcal{B}$ such that $n_{A_i} = 0$ and $n_{B_j} = 0$ simultaneously, the general log-odds problem,

$$\inf_{\boldsymbol{\alpha}\in\triangle^{|\Omega|}} E_{\text{Dir}}(g|\nu\boldsymbol{\alpha}+\boldsymbol{n}),$$

where,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

has the following interpretation at the boundary of $\triangle^{|\Omega|} \ni \boldsymbol{\alpha}$: the lower or upper bounds of $-\infty$ or $+\infty$ for the posterior expectation naturally occur as optima. It is an indication that the inference varies immensely over the set of Dirichlet priors of the IDM.

On the other hand, Theorem B.5.1 shows that the optimum may fail to exist as a limit point of an optimisation path when some sets in both the numerator and denominator of $g(\boldsymbol{\theta})$ have zero cell counts. This case is pathological in the sense that nothing can be done beyond interpreting the resulting inference as vacuous (i.e. that the conclusion of the posterior inference is simply that the posterior expectation is trivially between $-\infty$ and $+\infty$).

From this discussion, we can see that coherence of this inference breaks down when the data set is sparse as the mathematical justification for approximating with bounded random variables fails. On the other hand, the interpretation of the optimisation result as a form of global sensitivity analysis does not break down and is perfectly well defined. Hence, our posterior inference of the general log-odds demonstrates a divergence between the two statistical methodologies.

For the rest of this chapter, we will use the sensitivity analysis methodology to interpret any log-odds inference under the IDM and data sparsity.

3.4 Inference for log-odds under IDM with sparse observations

3.4.1 Behaviour of the posterior inference of the simple log odds under the IDM and sparse observations

We consider the optimisation of a single log odds, say,

$$\log \frac{\theta_i}{1 - \theta_1 - \ldots - \theta_{m-1}}.$$

We consider the unboundedness properties of the Dirichlet expected log odds along with the affine geometrical interpretation of a posterior update as being a translation of fixed

simplex around a larger simplex.

Example 3.4.1: Suppose that $m = 3$. Consider the various positions of the translated simplex relative to the larger simplex of possible (Dirichlet) natural parameters at the slice $s = \nu + n$ under different kinds of observations.

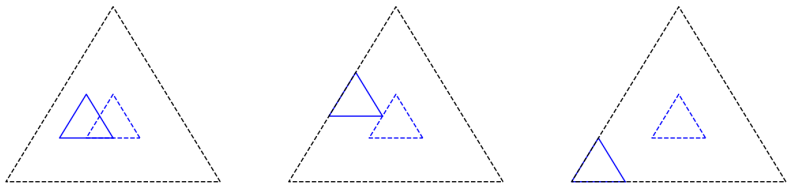

Figure 3.4: For $\nu = 2$, each subplot is associated with updating with different observation vectors totaling $n = 6$ observations. Left to right: $(n_1, n_2, n_3) = (1, 2, 3), (0, 3, 3), (0, 6, 0)$. The prior set of distributions with $\nu = 2$ (dashed blue) is translated by (n_1, n_2) to obtain the posterior set of distributions of Dirichlet natural parameters (solid blue.) The possible posterior Dirichlet natural parameters is the scaled simplex $(\nu + n)\overline{\blacktriangle^2}$ (dashed black.)

From Figure 3.4, when observations contains zero counts, the posterior set of distributions will always contain distributions whose natural parameters are on the boundary of the larger, ambient simplex (dashed black).

■

Example 3.4.2: Let us consider what happens to the mean of a simple log-odds under a posterior Dirichlet distribution when some cells have zero observations. These cases correspond to the middle and right panels in Figure 3.4. For three categories, when under the natural parametrisation $\boldsymbol{\theta} \sim \text{Dirichlet}((\nu + n, \nu\alpha_1 + n_1, \nu\alpha_2 + n_2))$,

$$\mu_i = E_{\text{Dir}}\left(\log \frac{\theta_i}{1 - \theta_1 - \theta_2} \,\middle|\, \boldsymbol{n}, \boldsymbol{\alpha}, \nu\right) = \psi(\nu\alpha_i + n_i) - \psi(\nu + n - \nu\alpha_1 - n_1 - \nu\alpha_2 - n_2).$$

for $i = 1, 2$. The middle panel corresponds the dataset $(n_1, n_2, n_3) = (0, n_2, n_3)$ where $n_2, n_3 > 0$. We observe that, for vertices $(0, \nu + n_2, n_3)$ and $(0, n_2, \nu + n_3)$, where $\alpha_2 = 1$

and $\alpha_3 = 1$ respectively,

$$(\mu_1, \mu_2)(0, \nu + n_2, n_3) = (\psi(0^+) - \psi(n - n_2), \psi(n_2) - \psi(n - n_2)),$$
$$(\mu_1, \mu_2)(0, n_2, \nu + n_3) = (\psi(0^+) - \psi(\nu + n - n_2), \psi(n_2) - \psi(\nu + n - n_2)),$$

where the right limit at 0 of the digamma function $\psi(0^+)$ tends to infinity. Notice that these two vertices represent endpoints of the side of the blue solid simplex that coincides with the boundary larger black simplex.

Similarly, the right panel corresponds to the dataset $(0, n_2, 0)$ for $n_2 = n > 0$. For the vertex $(0, \nu + n_2, 0)$,

$$(\mu_1, \mu_2)(0, \nu + n_2, 0) = (\psi(0^+) - \psi(0^+), \psi(\nu + n_2) - \psi(0^+)).$$

The indeterminate form $\psi(0^+) - \psi(0^+)$ in μ_1 arises from $(0, \nu + n_2, 0)$ coinciding with the vertex of the black simplex. This is qualitatively different from the terms that involve only one right limit of the digamma function at zero, which is merely unbounded.

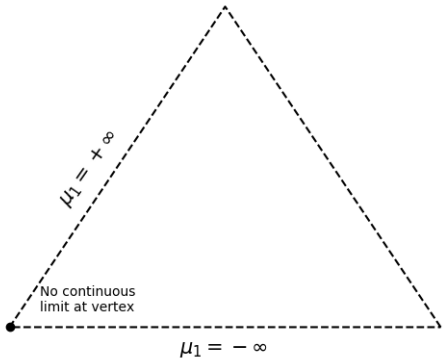

Figure 3.5: The ambient simplex $(\nu + n)\overline{\blacktriangle}^2$ with $(n_1, n_2, n_3) = (0, n_2, 0)$. The posterior expected log odds μ_1 takes values $+\infty$ and $-\infty$ on the left and bottom edges, respectively, and does not have a continuous limit at the vertex of these two edges.

The middle and right panels demonstrate qualitatively different boundary effects of the posterior natural parameter space upon log-odds inference depending on whether or not

the boundary contains vertices or not. This leads to the optimisation of the posterior expectation being also qualitatively very different amongst the three cases of boundaries in Figure 3.4.

■

3.4.2 Behaviour and solutions to optimisation problem of the posterior inference of the general log odds under the IDM and sparse observations

Let us now generalise the observations from the case with the simple log odds. The observations $\boldsymbol{n}$ can only affect the expected log-odds function only through the elements that appear explicitly in the linear combination of digamma functions,

$$\boldsymbol{\alpha} \longmapsto \sum_{i=1}^{r} \psi(\nu\alpha_{A_i} + n_{A_i}) - \sum_{j=1}^{s} \psi(\nu\alpha_{B_j} + n_{B_j}).$$

We can deduce the behaviour of this function through the behaviour of the difference between two digamma functions as follows.

Lemma 3.4.1: When $n_j > 0$,

$$\boldsymbol{\alpha} \longmapsto \psi(\nu\alpha_i) - \psi(\nu\alpha_j + n_j),$$

has a minimum of $-\infty$ as α_i tends to 0 on $\overline{\triangle^m} \ni \boldsymbol{\alpha}$. When $n_i, n_j > 0$, then the two terms in,

$$\psi(\nu\alpha_i + n_i) - \psi(\nu\alpha_j + n_j),$$

are bounded over $\boldsymbol{\alpha} \in \overline{\triangle^m}$.

Proof: the first result is due to the fact that ψ is an unbounded and increasing function on $\mathbb{R}^+$. The last result is also clear as $\psi(x)$ is finite when x is finite.

□

In particular, the appropriate categories having zero observations causes the entire posterior expected log odds to go to infinity in one or both directions.

We note that the optima of the expectation of the general log-odds remain infinite as long as at least one event has zero counts. In particular, we have the following.

Theorem 3.4.1:

- Suppose that $n_{A_1} = 0$ and $n_{A_i}, n_{B_j} > 0$ for $i \neq 1$. Then, the lower expectation of the general log-odds is $-\infty$.
- Suppose that $n_{B_1} = 0$ and $n_{A_i}, n_{B_j} > 0$ for $j \neq 1$. Then, the upper expectation of the general log-odds is $+\infty$.

Proof: For the first case, the expectation to be minimised is,

$$\alpha \in \overline{\triangle^m} \longmapsto \psi(\nu\alpha_{A_1}) + \sum_{i=2}^{r} \psi(\nu\alpha_{A_i} + n_{A_i}) - \sum_{j=1}^{s} \psi(\nu\alpha_{B_j} + n_{B_j}).$$

We need only demonstrate that $-\infty$ is attainable and so must be the minimum. Indeed, any solution on the face $\left\{\alpha : \ \alpha_{A_1} = 0, \ \bigwedge_{i=2}^{r} \alpha_{A_i} > 0 \bigwedge_{j=1}^{s} \alpha_{B_j} > 0\right\}$ will suffice.

The case for $n_{B_1} = 0$ is analogous, with $\left\{\alpha : \ \alpha_{B_1} = 0, \ \bigwedge_{i=1}^{r} \alpha_{A_i} > 0 \bigwedge_{j=2}^{s} \alpha_{B_j} > 0\right\}$ attaining $+\infty$.

□

Finally, we remark upon the indeterminate forms that appear in certain expressions for the posterior expectation. For example, in Example 3.4.2, the expression for the posterior expectation of the log-odds was,

$$\psi(\nu\alpha_i + n_i) - \psi(\nu + n - \nu\alpha_1 - n_1 - \nu\alpha_2 - n_2).$$

for $i = 1, 2$, such that with $i = 1$, $n_1 = n_2 = 0$, the indeterminate form $\psi(0^+) - \psi(0^+)$ occurs as α_1 and α_2 approach 0. This is because the limit is not well-defined: the indeterminate form can take any real value depending on the path taken for α_1, α_2 to approach 0. This

simply means that the posterior expectation $E_{\text{Dir}}\left(\log \frac{\theta_i}{1-\theta_1-\theta_2} \mid \boldsymbol{n}, \boldsymbol{\alpha}, \nu\right)$ cannot be assigned the value returned by this expression, and must be directly evaluated. Consequently, if this case occurs when computing this type of imprecise expectation, we minimise the Lebesgue integral directly.

3.4.3 Effects of cell counts on imprecision of posterior log-odds inference under the IDM

We have seen that zero sample counts in categories may cause the IDM's lower and upper expectations to be unbounded, and the inference to become noninformative. On the other hand, as long as the categories involved in the log-odds contain at least one observation, the expectation becomes finite. As next step, let us explore how observing a single observation in its event decreases the imprecision of the posterior inference. Suppose that $n_A = 1$ such that,

$$E[\log \theta_A | \boldsymbol{n}, \nu, \boldsymbol{\alpha}] = \psi(\nu\alpha_A + 1) - \psi(\nu + n).$$

Then the minimum is attained at,

$$\psi(1) - \psi(\nu + n),$$

and the maximum is attained at,

$$\psi(\nu + 1) - \psi(\nu + n).$$

In terms of gain of precision, one observation in A causes the lower expectation to jump from $-\infty$ to the finite value $\psi(1) - \psi(\nu + n)$: thus ν and n also controls the location of the lower expectation when transitioning away from vacuity. The imprecision decreases from ∞ to $\psi(\nu+1) - \psi(1)$ with a single observation $n_A = 1$, and higher values of ν increases this imprecision value. Interestingly, the total number of observations n does not contribute to this phenomenon.

As expected, as more counts are accumulated by increasing Δn, the imprecision of the log-probability inference also decreases.

Proposition 3.4.1: for $\boldsymbol{n}$ and $\Delta\boldsymbol{n}$ such that $n_A \geq 1$ and $\Delta n_A = (\Delta\boldsymbol{n})_A \geq 0$,

$$\overline{P}[\log \theta_A | \nu\boldsymbol{\alpha} + \boldsymbol{n}] \geq \overline{P}[\log \theta_A | \nu\boldsymbol{\alpha} + \boldsymbol{n} + \Delta\boldsymbol{n}].$$

Proof: See Proposition B.4.1.

□

Due to a triangular inequality of the imprecision of a sum (see Lemma B.4.1,) the imprecision of a general log-odds is bounded above by the imprecision of the sum of its components of log-probabilities.

Theorem 3.4.2: For finite r, s, the imprecision (induced by a coherent conjugate pair of lower and upper expectations) of the general log odds is bounded by the imprecisions of its component log-probabilities:

$$\overline{P}\left[\log \frac{\theta_{A_1} \dots \theta_{A_r}}{\theta_{B_1} \dots \theta_{B_s}}\right] \leq \sum_{i=1}^{r} \overline{P}[\log \theta_{A_i}] + \sum_{j=1}^{s} \overline{P}[\log \theta_{B_j}].$$

Proof: Apply Lemma B.4.1.

□

3.5 Illustrative numerical examples

3.5.1 Setting for numerical optimisation

Optimisations are run using the `spg` function of the `R` library, `BB` (Varadhan and Gilbert [79]). This is a variant of constrained gradient descent that projects the solution back onto the domain if the gradient steps out of it (see Birgin et al. [16]). This is particularly useful for optimisations with domains that involve closed sets in Euclidean space where the optimum may exist in a subset that is in a lower dimensional subspace (such as the boundaries of the closed simplex that parametrises the posterior set of distributions of the IDM.)

For initialisation of the optimisation on $\triangle^m$, we instantiate by sampling $z \sim N(0, I_k)$, taking the absolute value of each element and normalising the vector:

$$\frac{(|z_1(\omega)|, \ldots, |z_k(\omega)|)}{\sum_{i=1}^{m} |z_i(\omega)|}. \tag{3.8}$$

Finally, because we are interested specifically in effects due to the patterns of sparsity in the data, we fix the IDM parameter $\nu = 2$ for the experiments.

3.5.2 Dataset examples

Example 3.5.1: Hockey goals data: log-odds inference

Goals Scored / Goals Given Up	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	1	2	1	0	1	1	0	0	0	1
2	0	1	3	2	0	3	1	0	0	0	1	0	0
3	0	1	0	2	3	2	6	3	0	1	0	0	0
4	2	2	1	3	3	2	2	0	0	1	0	0	0
5	0	1	2	3	2	1	2	0	0	1	0	0	0
6	0	1	1	2	1	1	0	1	0	0	0	0	0
7	0	1	1	0	0	1	2	0	0	0	0	0	0
8	1	0	0	0	1	1	0	0	0	0	0	0	0

This dataset was analysed by Dong and Simonoff [37], whose interest was in constructing smoothed estimators of cell probabilities θ^*_{ij}, for the i-j-th cell in the table. The data consists of observations of a total of '80 games played by the Pittsburgh Penguins of the National Hockey League during the 1991-1992 season' [37]. The i-j-th cell represents the count of the games where the team gave up i goals and scored j goals. The extent of their analysis is a heat plot of smooth counts $n\theta^*_{ij}$ (where θ^*_{ij} is their smoothed cell probability estimator:) in particular, the authors remarked that, '...typical for a sparse table like this, drawing any conclusions from the table is difficult, past ... that in most games the team both scored and gave up between 1 and 7 goals' [37]. They observed that their smoothing highlights regions of negative correlation such as high smoothed counts in cells 2-5, 3-5, 3-6, 5-2, and 5-3.

We will first consider this hypothesis by computing the imprecise expectation of the odds ratio of representing stochastic dominance. Namely,

$$g(\boldsymbol{\theta}) = \log \frac{P(\text{GGU} > \text{GS})}{1 - P(\text{GGU} > \text{GS})}(\boldsymbol{\theta}) = \log \sum_{i>j} \theta_{ij} - \log \sum_{i \leq j} \theta_{ij},$$

where GGU and GS are the counts of goals given up and scored, respectively. The optimisation objective function for the imprecise expectations under the IDM is the following posterior Dirichlet expectation,

$$\boldsymbol{\alpha} \in \overline{\triangle^m} \longmapsto E\left[\log \frac{P(\text{GGU} > \text{GS})}{1 - P(\text{GGU} > \text{GS})}(\boldsymbol{\theta}) \,\middle|\, \nu\alpha + \boldsymbol{n}\right] = \psi\left(\sum_{i>j} \nu\alpha_{ij} + n_{ij}\right) - \psi\left(\sum_{i \leq j} \nu\alpha_{ij} + n_{ij}\right).$$

Per Walley's suggestion [80], we choose $\nu = 2$ and, as with the illustrative examples, perform 100 repeats of the optimisation with the random initialisation (3.8).

Minimum of $\underline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\underline{P}(g\|\nu, \boldsymbol{n})$
-0.452	-0.452

Minimum of $\overline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\overline{P}(g\|\nu, \boldsymbol{n})$
-0.349	-0.349

All is well in terms of our earlier analysis of sparsity: both optimisation problems of the lower and upper expectations are convergent (as shown by the repeated testing) to a finite number. This is due to there being at least one cell of the event $\{\text{GGU} > GS\}$ having at least one observation such that $\sum_{i>j} n_{ij} > 0$, and similarly for $\sum_{i \leq j} n_{ij} > 0$. In fact,

both the lower and upper expectations are negative, suggesting agreement amongst the prior distributions that the 'correlation' is indeed negative. Unlike the lower expectations, amongst the finite upper expectations, there is a variety of signs being taken by the upper expectation values.

■

Example 3.5.2: Hockey goals data: log odds ratio inference

Now consider instead the log odds ratio,

$$\begin{aligned} g(\boldsymbol{\theta}) &= \log \frac{P(\text{GGU} > \text{GS} \cap \text{GS} \geq 2)P(\text{GGU} \leq \text{GS} \cap \text{GS} < 2)}{P(\text{GGU} > \text{GS} \cap \text{GS} < 2)P(\text{GGU} \leq \text{GS} \cap \text{GS} \geq 2)}(\boldsymbol{\theta}) \\ &= \log \frac{P(\text{GGU} > \text{GS}|\text{GS} \geq 2)/P(\text{GGU} \leq \text{GS}|\text{GS} \geq 2)}{P(\text{GGU} > \text{GS}|\text{GS} < 2)/P(\text{GGU} \leq \text{GS}|\text{GS} < 2)}(\boldsymbol{\theta}). \end{aligned}$$

In other words, we infer on the comparison between the odds of losing a game conditional on the team scoring at least or less than 2 goals. Again, we check for sensitivity to random instantiations by reporting the minimum and maximum of lower and upper expectations over 100 repeats:

Minimum of $\underline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\underline{P}(g\|\nu, \boldsymbol{n})$
-3.30	-3.30

Minimum of $\overline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\overline{P}(g\|\nu, \boldsymbol{n})$
-2.93	-2.93

For all $\boldsymbol{\alpha}$ parametrising each Dirichlet distribution in the IDM set of priors $\{\text{Dirichlet}(\cdot|\nu = 2, \boldsymbol{\alpha}) : \ \boldsymbol{\alpha} \in \overline{\triangle^{8\times 12}}\}$,

$$E_{P_{\text{Dir}}}\left(\log \frac{P(\text{GGU} > \text{GS}|\text{GS} \geq 2)/P(\text{GGU} \leq \text{GS}|\text{GS} \geq 2)}{P(\text{GGU} > \text{GS}|\text{GS} < 2)/P(\text{GGU} \leq \text{GS}|\text{GS} < 2)}(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{n}, \nu\boldsymbol{\alpha}\right) < 0,$$

or,

$$\begin{aligned} &E_{P_{\text{Dir}}}\left(\log P(\text{GGU} > \text{GS}|\text{GS} \geq 2)/P(\text{GGU} \leq \text{GS}|\text{GS} \geq 2)(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{n}, \nu\boldsymbol{\alpha}\right) \\ &< E_{P_{\text{Dir}}}\left(\log P(\text{GGU} > \text{GS}|\text{GS} < 2)/P(\text{GGU} \leq \text{GS}|\text{GS} < 2)(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{n}, \nu\boldsymbol{\alpha}\right). \end{aligned}$$

In accordance with the observed data, the posterior distributions in the set of distributions agree that the expected odds of them losing (goals given up (GGU) greater than goals scored (GS)) conditional on them scoring at least 2 goals is less than the same odds conditional on scoring less than 2 goals, which is to be expected.

■

Example 3.5.3: Hockey goals data: log odds inference with finer conditioning on rare events

Suppose that one was interested in comparing the odds of them losing under the regimes when the number of goals given up (GGU) and goals scored (GS) are low (that is, when both teams' scores are low) and when both are high. This is also numerically interesting because these cases are also part of the sparse parts of the dataset (the upper-left and lower-right portions of it).

One way to model that is to split the values of the GGU and GS variables such that the conditioning is as follows. We condition on both GGU and GS being low, which we define as when they are in $\{0, 1\}$ and $\{0, 1\}$ respectively, and when both are high, when they are in $\{6, 7, 8\}$ and $\{8, 9, 10, 11, 12\}$ respectively. The conditioning events are sparse: they are chosen such that the variable is ordinally smaller than the smallest value whose cell count is greater than or equal to 3 occur (for example, GGU $\in \{0, 1\}$ was chosen because GGU $= 2$ is the smallest ordinal value whose row starts containing cell counts of at least 3.)

The log-odds can be written as,

$$g(\boldsymbol{\theta}) = \log \frac{P(\text{GGU}{\geq}\text{GS} \cap \text{GGU}{\in}\{0,1\} \cap \text{GS}{\in}\{0,1\})P(\text{GGU}{<}\text{GS} \cap \text{GGU}{\in}\{6,7,8\} \cap \text{GS}{\in}\{8,9,10,11,12\})}{P(\text{GGU}{\geq}\text{GS} \cap \text{GGU}{\in}\{6,7,8\} \cap \text{GS}{\in}\{8,9,10,11,12\})P(\text{GGU}{<}\text{GS} \cap \text{GGU}{\in}\{0,1\} \cap \text{GS}{\in}\{0,1\})}(\boldsymbol{\theta}). \tag{3.9}$$

In terms of conditional odds-ratios,

$$\log \frac{P(\text{GGU}{\geq}\text{GS} \mid \text{GGU}{\in}\{0,1\} \cap \text{GS}{\in}\{0,1\})/P(\text{GGU}{\geq}\text{GS} \mid \text{GGU}{\in}\{6,7,8\} \cap \text{GS}{\in}\{8,9,10,11,12\})}{P(\text{GGU}{<}\text{GS} \mid \text{GGU}{\in}\{0,1\} \cap \text{GS}{\in}\{0,1\})/P(\text{GGU}{<}\text{GS} \mid \text{GGU}{\in}\{6,7,8\} \cap \text{GS}{\in}\{8,9,10,11,12\})}(\boldsymbol{\theta}).$$

Labelling the log-odds in (3.9), $\log \theta_A\theta_B/\theta_C\theta_D$, its posterior expectation with respect to one set of distributions element of the IDM is,

$$\psi(\nu\alpha_A) + \psi(\nu\alpha_B) - \psi(\nu\alpha_C) - \psi(\nu\alpha_D),$$

(such that $n_A, n_B, n_C, n_D = 0$.) From this, $-\infty$ is the minimum that can be found on the face of $\{\alpha : \alpha_A = 0 \vee \alpha_B = 0\}$. The maximum ∞ can be found on $\{\alpha : \alpha_C = 0 \vee \alpha_D = 0\}$. Again, we check for sensitivity by taking the minimum and maximum over 100 random initialisations:

Minimum of $\underline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\underline{P}(g\|\nu, \boldsymbol{n})$
$-\infty$	$-\infty$

Minimum of $\overline{P}(g\|\nu, \boldsymbol{n})$	Maximum of $\overline{P}(g\|\nu, \boldsymbol{n})$
$+\infty$	$+\infty$

Compared with Example 3.5.2, by changing the conditioning events to a collection that does not partition the indices of the table and are rare, the imprecision becomes effectively infinite and the inference is consequently vacuous.

■

Example 3.5.4: Independence test on modified Hockey Data

We will now draw inference on the independence between the variables GGU and GS. Under the multinomial likelihood, these two random variables are independent (relative to the categories $I := \{0, 1, \ldots, 8\} \times \{0, 1, \ldots, 12\}$) iff

$$\forall (i, j) \in I : \ \log \frac{\theta_{i.}\theta_{.j}}{\theta_{ij}} = 0. \tag{3.10}$$

We will be computing the imprecise expectation of these $9 \times 13 = 117$ log odds and, as with our previous examples, we will perform repetitions of each whilst varying only the initialisation of each optimisation run. Under a posterior Dirichlet$(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ distribution, its expectation is,

$$E\left[\log \frac{\theta_{i.}\theta_{.j}}{\theta_{ij}} \,\middle|\, \nu\boldsymbol{\alpha} + \boldsymbol{n}\right] = \psi(\nu\alpha_{i.} + n_{i.}) + \psi(\nu\alpha_{.j} + n_{.j}) - \psi(\nu\alpha_{ij} + n_{ij}) - \psi(\nu + n). \tag{3.11}$$

In general, the minimum of the expected log-odds (3.11) is bounded away from $-\infty$ since all columns and rows sum to at least 1, so that it cannot be achieved by any $\boldsymbol{\alpha} \in \overline{\triangle^m}$. The maximum is also similarly bounded away from ∞ when $n_{ij} > 0$, but is ∞ when $n_{ij} = 0$ and the optimum tends to some maximiser satisfying $\alpha_{ij} = 0$. As a general remark, we

expect this as the statistic itself $\log \theta_{i\cdot}\theta_{\cdot j}/\theta_{ij}$ for $\boldsymbol{\theta} \in \overline{\triangle^m}$ is unbounded above (as θ_{ij} tends to 0) but bounded below away from $-\infty$ as $\theta_{i\cdot}, \theta_{\cdot j} \geq \theta_{ij}$ prevents the numerator of the ratio from going to zero when $\theta_{ij} > 0$.

Notice that because the column GS = 11 has no counts, it can be seen that,

$$\forall i \in \{0, 1, \ldots, 8\} : \; E\left[\log \frac{\theta_{i\cdot}\theta_{\cdot 11}}{\theta_{i11}} \;\middle|\; \nu\boldsymbol{\alpha} + \boldsymbol{n}\right] = \psi(\nu\alpha_{i\cdot} + n_{i\cdot}) + \psi(\nu\alpha_{\cdot 11}) - \psi(\nu\alpha_{i11}) - \psi(\nu + n),$$

which is minimised and maximised to $-\infty$ and ∞ respectively, so it is not particularly interesting as far as satisfying (3.10) goes.

Instead, we will modify the hockey data slightly by adding a single unit count to the column $GS = 11$. We add a single count to the cell (0,11). The modified table is as follows,

Goals Scored Goals Given Up	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	1	0	0	1	0
1	0	0	0	1	2	1	0	1	1	0	0	0	1
2	0	1	3	2	0	3	1	0	0	0	1	0	0
3	0	1	0	2	3	2	6	3	0	1	0	0	0
4	2	2	1	3	3	2	2	0	0	1	0	0	0
5	0	1	2	3	2	1	2	0	0	1	0	0	0
6	0	1	1	2	1	1	0	1	0	0	0	0	0
7	0	1	1	0	0	1	2	0	0	0	0	0	0
8	1	0	0	0	1	1	0	0	0	0	0	0	0

Notice that none of the rows and columns sums to zero now.

The lower and upper expectations of this problem is given in Table 3.1. (Based on 100 random initialisations, we found that the results were not sensitive to the random initialisations. The output of these tests have been omitted for brevity.)

	0	1	2	3	4	5	6	7	8
0	[-2.07, $+\infty$]	[-1.43, $+\infty$]	[-1.32, $+\infty$]	[-0.90, $+\infty$]	[-0.97, $+\infty$]	[-0.97, $+\infty$]	[-0.90, $+\infty$]	[-1.70, $+\infty$]	[-3.04, -1.99]
1	[-1.18, $+\infty$]	[-0.55, $+\infty$]	[-0.44, $+\infty$]	[-0.52, 0.83]	[-0.92, -0.25]	[-0.59, 0.74]	[-0.02, $+\infty$]	[-1.32, -0.09]	[-1.93, -0.70]
2	[-0.80, $+\infty$]	[-0.66, 0.65]	[-1.14, -0.73]	[-0.46, 0.21]	[0.29, $+\infty$]	[-0.79, -0.36]	[-0.13, 1.21]	[-0.43, $+\infty$]	[-1.05, $+\infty$]
3	[-0.35, $+\infty$]	[-0.22, 1.16]	[0.38, $+\infty$]	[-0.02, 0.70]	[-0.34, 0.12]	[-0.09, 0.62]	[-0.78, -0.58]	[-1.07, -0.60]	[-0.60, $+\infty$]
4	[-1.30, -0.58]	[-0.66, 0.04]	[-0.22, 1.15]	[-0.38, 0.07]	[-0.45, 0.00]	[-0.20, 0.50]	[-0.13, 0.57]	[-0.10, $+\infty$]	[-0.71, $+\infty$]
5	[-0.72, $+\infty$]	[-0.59, 0.74]	[-0.81, -0.14]	[-0.64, -0.20]	[-0.46, 0.21]	[-0.12, 1.21]	[-0.39, 0.29]	[-0.35, $+\infty$]	[-0.97, $+\infty$]
6	[-1.18, $+\infty$]	[-1.05, 0.19]	[-0.94, 0.32]	[-0.85, -0.16]	[-0.59, 0.74]	[-0.59, 0.74]	[-0.02, $+\infty$]	[-1.32, -0.09]	[-1.43, $+\infty$]
7	[-1.45, $+\infty$]	[-1.32, -0.09]	[-1.21, 0.05]	[-0.28, $+\infty$]	[-0.35, $+\infty$]	[-0.85, 0.47]	[-1.12, -0.43]	[-1.09, $+\infty$]	[-1.70, $+\infty$]
8	[-2.33, -1.32]	[-1.18, $+\infty$]	[-1.07, $+\infty$]	[-0.65, $+\infty$]	[-1.22, 0.11]	[-1.22, 0.11]	[-0.65, $+\infty$]	[-1.45, $+\infty$]	[-2.07, $+\infty$]

	9	10	11	12
0	[-2.07, $+\infty$]	[-2.65, $+\infty$]	[-3.99, -2.42]	[-2.65, $+\infty$]
1	[-1.18, $+\infty$]	[-1.77, $+\infty$]	[-1.77, $+\infty$]	[-2.54, -1.04]
2	[-0.80, $+\infty$]	[-2.06, -0.56]	[-1.38, $+\infty$]	[-1.38, $+\infty$]
3	[-0.85, 0.53]	[-0.94, $+\infty$]	[-0.94, $+\infty$]	[-0.94, $+\infty$]
4	[-0.96, 0.41]	[-1.05, $+\infty$]	[-1.05, $+\infty$]	[-1.05, $+\infty$]
5	[-1.22, 0.11]	[-1.30, $+\infty$]	[-1.30, $+\infty$]	[-1.30, $+\infty$]
6	[-1.18, $+\infty$]	[-1.77, $+\infty$]	[-1.77, $+\infty$]	[-1.77, $+\infty$]
7	[-1.45, $+\infty$]	[-2.04, $+\infty$]	[-2.04, $+\infty$]	[-2.04, $+\infty$]
8	[-1.82, $+\infty$]	[-2.40, $+\infty$]	[-2.40, $+\infty$]	[-2.40, $+\infty$]

Table 3.1: Lower and upper expectations $[\underline{E}(g|\nu, \boldsymbol{n}), \overline{E}(g|\nu, \boldsymbol{n})]$ of the independence log-odds statistic g under the modified hockey dataset.

Because there are some cells whose imprecise intervals contain 0, while others do not, there is no agreement amongst the elements about the independence statement (3.10). This leads us to conclude that, under this modified hockey dataset, this particular inference is sensitive to the choice of prior at which the posterior expectation is evaluated, to the point that posterior inference from the different priors do not even agree on the sign of the posterior expectation of the log-odds in (3.10) for some cells. Indeed, the pair of lower and upper expectations $(\underline{E}(g|\nu, \boldsymbol{n}), \overline{E}(g|\nu, \boldsymbol{n}))$ of each cell fall into one of the following cases in terms of their signs.

1. $(-, -)$: all prior Dirichlet distributions produce posterior expectations of the independence statistic (3.11) which are negative. Conclusively, for such cells $i - j$, the priors all agree on the statement $E_P[\log \theta_{i\cdot}\theta_{\cdot j}/\theta_{ij}] < 0$.

2. $(-, +)$ and $(-, \infty)$: some prior Dirichlet distributions produce posterior expectations of the independence statistic (3.11) which are negative, while some produce positive values. Consequently, there is no consensus on the sign of $E_P[\log \theta_{i\cdot}\theta_{\cdot j}/\theta_{ij}]$. In particular, the null value of interest 0, is in their imprecise interval.

3. $(+, \infty)$: all prior Dirichlet distributions produce posterior expectations of the independence statistic (3.11) which are positive. Conclusively, for such cells $i - j$, the priors all agree on the statement $E_P[\log \theta_{i\cdot}\theta_{\cdot j}/\theta_{ij}] > 0$.

For some cells, the upper expectation is unbounded. For the cells which are otherwise not, the upper expectation seemss to also be invariant against the random initialisations (3.8). In fact, the cells that correspond to upper expectations that are bounded are exactly the ones with at least one observation.

■

3.6 Concluding remarks

Our contributions in this chapters can be summarised as follows. We construct Example 3.1.1 to motivate the material presented in this chapter. Example 3.2.2 and Theorem 3.2.2 demonstrate the geometry of IDM posterior update, pictorially demonstrated by Figure 3.3. Theorem 3.3.1 yields the evaluation of the IDM coherent lower expectation of the unbounded general log-odds. Example 3.3.1 demonstrates that this extension fails in certain

cases when the data is sparse. Subsequently, Corollary 3.3.2 argues that global sensitivity analysis is a more suitable interpretation of the IDM log-odds problem in sparse data cases. Using Examples 3.4.1, 3.4.2, we demonstrate the behaviour of the posterior IDM optimisation of the log-odds statistics at the boundary of the natural parameter space. We briefly explore the amount of precision gained for one unit of observations: Theorem 3.4.2 provides a triangular inequality bound on its imprecision. Finally, inference on real dataset were demonstrated in Examples 3.5.1, 3.5.2, 3.5.3 and 3.5.4.

We observed that for some sparse multinomial observations, the likelihood can be flat in some directions and therefore Bayesian posterior inference will be to some extent dictated by (and therefore sensitive to) the choice of prior distribution. Imprecise models such as the IDM allow us to preserve a generalisation of the usual coherence that forms a pillar of Bayesian inference (see Chapter 2) and a global sensitivity analysis interpretation allows us to consider the effect of sparsity and the use of multiple (Dirichlet) prior distributions on the inference by considering the so-called imprecise inference that consists of lower and upper bounds of posterior expectations.

Following Bickis [14], the posterior update can be geometrically represented by a translation of the set of (natural parameters of) individual probabilistic models, and sparsity is readily interpreted as a simplex being translated along the face of a larger simplex, motivating the need to consider the extended Dirichlet family. We focused our analysis on the properties of posterior imprecise expectation of a general form of log-odds and derived its properties as an optimisation of sums and differences of digamma functions and the latter's asymptotic properties were heavily used in the analysis. Finally, synthetic and data examples were used for numerical illustration for the model.

Like the Dirichlet-Multinomial model, the IDM also depends on the multinomial data only through the cell counts. When the counts of interest are zero, because the IDM includes some degenerate Dirichlet distributions of lower dimensions, the posterior log-odds expectation value will gravitate towards $\pm\infty$ and our analysis and examples readily corroborated this.

When sparsity is not so severe, the IDM does provide a non-vacuous global sensitivity analysis. For example, in the Hockey goals data examples 3.5.1 and 3.5.2, the counts involved are all non-zero, and the IDM produces finite (but different) posterior lower and upper expectations, yielding a meaningful inference by measuring the discrepancies amongst the

priors through the non-zero imprecision even though some of the component cells themselves have zero count.

Aggregation plays the rôle of making inference less vacuous. For example, comparing Examples 3.5.2 and 3.5.3, when in the former we condition on events which have non-zero counts and on events which are finer and are of zero counts in the latter, the latter produces vacuous lower and upper expectations. In the modified Hockey data example, Example 3.5.4, even when the rows and columns of the dataset have positive marginal counts, the fact that the independence statistic depend on each individual cell count in the denominator that can still be zero means that, again, in this case the IDM can provide only (upper) vacuous inference.

More generally, the modified hockey example brings up some methodological problems in hypothesis testing and estimation using the imprecise framework. In the precise Bayesian and frequentist settings, a useful interpretive tool to account for variations of point estimates due to the randomness from the distribution is the construction of intervals such as credible and confidence intervals (incoherence of the confidence interval not withstanding, see Chapter 7 of Walley [81]). As we have discussed in Section 2.4.4, the interval of a pair of imprecise expectations, $[\underline{E}(f(\boldsymbol{\theta})), \overline{E}(f(\boldsymbol{\theta}))]$, does not have the same interpretation as the credible or confidence interval, as it is merely a collection of point estimates from a different prior.

To truly extend the usual statistical methodology of interval estimation to the imprecise realm, the variation of each prior must also somehow be aggregated to result in a summary of, say, the imprecision of credible intervals over a set of priors, whichever form this sensitivity analysis may manifest. To our knowledge, this has not been explored in the literature and, to our example of testing for the independence of the underlying data generating process of the modified hockey example, it seems inadequate to conclude anything about the independence hypothesis by only computing a set of point estimates. An attempt to address this is made in Chapter 4 where we explore summarising posterior quantiles due to a set of prior distributions, and give interpretation to lower and upper quantiles at different percentiles as a type of 'imprecise interval' over a set of distributions.

Chapter 4

Imprecise quantile functions and interval-valued statistics in the imprecise setting

In this chapter, we apply the imprecise probabilities methodology in the context of hypothesis testing and estimation of a univariate parameter. Work has been done on hypothesis testing using lower and upper probabilities (for example, Couso, Álvarez-Caballero and Sánchez [26] and Perolat, Couso, Loquin and Strauss [65]) but estimation still remains to be developed. One issue of estimation in the imprecise setting is that no generalisation of interval statistics such as confidence or posterior credible intervals have been proposed in the same way that imprecise expectations coherently generalise the precise point estimate using the expectation.

We are inspired by the work of Couso, Moral and Sánchez [30] to construct an *imprecise quantile function* over a set of distributions. Because the quantile operator is generally nonlinear, unlike the expectation operator, its optimisation is not guaranteed by the lower envelope theorem (Theorem 2.2.1) to be a tight bound over a convex set of distributions. To that issue, we show that, under certain regularity conditions, its optima over a convex set of distributions also occur over the extreme points of this set. We provide data examples of how such imprecise quantiles can be used and, more importantly, interpreted in testing and estimation of a univariate parameter in the context of contingency tables under the imprecise Dirichlet model (IDM, Chapter 2, due to Walley [81, 80]).

4.1 Quantiles and imprecision

The notion of a quantile plays a major rôle in statistical theory and practice. In hypothesis testing, it may be used to define the acceptance region of some prescribed confidence or credibility. In parameter estimation, one usually wishes to not only have a point estimate, but to also measure the variability of this estimate. This use of quantiles is important in Bayesian analysis as the whole posterior distribution can be reported in order to preserve all the derived information (Gelman et al. [41]): in particular, the set of all quantiles, being the inverse of the distribution's cumulative function, describes the whole posterior distribution.

On the other hand, the imprecise expectations, introduced in Chapter 2, are lacking for the purposes of statistical inference for the following reason. From the definition of imprecision, an imprecise expectation, $[\underline{E}_M(X), \overline{E}_M(X)]$ is to be interpreted as all *point estimates* of the mean coherent with your state of knowledge: it does not summarise the effects of the uncertainty or randomness of each distribution in M. Indeed, the typical tools to indicate this are intervals such as confidence and credible intervals, but, to the best of our knowledge, there have been no development of their imprecise analogues used for statistical inference.

There are some ways to try and reproduce interval statistics in the imprecise methodology. For example, Walley [81] justifies the lower and upper variances,

$$\underline{V}_M(X) := \inf\ \{\mathrm{Var}_P(X) :\ P \in M\},\ \overline{V}_M(X) := \sup\ \{\mathrm{Var}_P(X) :\ P \in M\},$$

in the same gamble-theoretic terms as the lower and upper expectations (see Chapter 2). Heuristically, one may combine them in a sensitivity analysis manner. For example, for some constant $c > 0$, the following imprecise interval,

$$\left[\underline{E}_M(X) - c\sqrt{\overline{V}_M(X)}, \overline{E}_M(X) + c\sqrt{\overline{V}_M(X)}\right],$$

will contain all the intervals of the form $[E_P(X) \pm c\sqrt{\mathrm{V}_P(X)}]$ over $P \in M$. However, it is unclear how to formally extend any concept of coherence to this. For example, we can demonstrate that this interval may exceed the domain of a bounded finite random variable. This raises the question of how avoidance of sure losses may be defined for such interval summaries.

Alternatively, we propose to compute lower and upper quantile functions directly over a set of distributions M. That is, for a fixed percentile p and random variable X, we compute the minimum $\underline{\mathrm{Q}}_M^{(p)}(X)$ and maximum $\overline{\mathrm{Q}}_M^{(p)}(X)$ of the quantile function as a function of the underlying distribution,

$$P \longmapsto Q_P^{(p)}(X) = \inf\ \{u :\ P(X \leq u) \geq p\}.$$

Like the imprecise expectation, these will be lower and upper bounds, $\underline{\mathrm{Q}}_M^{(p)}(X), \overline{\mathrm{Q}}_M^{(p)}(X)$ of the p-th quantile over M.

For statistical inference, an imprecise interval statistic can be constructed by considering $\alpha < \beta \in [0,1]$, $[\underline{\mathrm{Q}}_M^{(\alpha)}(X), \overline{\mathrm{Q}}_M^{(\beta)}(X)]$. This construction has several advantages. Firstly, statements about bounds on imprecise quantiles can be shown to correspond to statements about bounds of the lower and upper CDFs (Theorem 4.5.1), which are coherent quantities. However, quantiles themselves do not readily fit into the coherence framework introduced in Chapter 2. Nevertheless, we show that just as the CDF and quantiles are monotonically connected in the sense that,

$$P(X \leq q) \geq p \iff \mathrm{Q}_P^{(p)}(X) \leq q,$$

we show in Theorem 4.5.2 that a similar relation occurs for the imprecise quantiles and its associate coherent lower and upper probabilities. Secondly, it has a sensitivity analysis interpretation: it contains all the α-β intervals of every distribution in M. Third, consequently, $1-\beta-\alpha$ is in fact a lower bound of the coverage probability of such α-β intervals of each distribution in M (Theorem 4.5.3).

4.2 Literature: Imprecise Quantiles

Little research investigates the extension of the quantile to the imprecise realm. For our purposes, the most relevant one can be found in the work done by Couso, Moral and Sánchez [30], who extend the analysis of a median set of a single distribution to a collection of such sets over a set of distributions. Let us briefly go over their construction. Note that they construct the median as a *set* of values.

Definition 4.2.1: (Couso, Moral and Sánchez [30]) For a distribution P and a random

variable X, the *median* of X under P is the set of values:

$$\text{Me}_P(X) := \{x : \ P(X \geq x) \geq 0.5 \ \wedge \ P(X \leq x) \geq 0.5\}.$$

□

Definition 4.2.2: (Couso, Moral and Sánchez [30]) For M a set of distributions, define *the lower and upper median of* X, respectively, as,

$$\underline{\text{Me}}_M(X) := \inf_{P \in M} \ \inf \text{Me}_P(X),$$

and

$$\overline{\text{Me}}_M(X) := \sup_{P \in M} \ \sup \text{Me}_P(X).$$

□

4.3 Imprecise Quantile Functions

Our goal is to combine the concept of quantile with imprecision in a way that is in line with the frameworks already introduced in Chapter 2. In probability theory, a quantile can be defined as a generalised inverse of the CDF. It is therefore reasonable to generalise this to the imprecise case by taking the lower and upper cumulative distribution functions to directly leverage the machinery presented in Chapter 2. However, as we have alluded already, imprecise quantiles such as the imprecise median in Definition 4.2.2 by Couso, Moral and Sánchez may not be readily analysed using this machinery.

Consider that,

$$\begin{aligned}
&\inf_{P \in M} \inf \{x : \ P(X \leq x) \geq 0.5\} \geq q \\
&\Leftrightarrow \ \forall P \in M : \inf \{x : \ P(X \leq x) \geq 0.5\} \geq q \\
&\Leftrightarrow \ \forall P \in M : P(X \leq q) \leq 0.5 \\
&\Leftrightarrow \ \overline{P}(X \leq q) \leq 0.5.
\end{aligned}$$

However,

$$\inf_{P\in M} \inf\{x:\ P(X\leq x)\geq 0.5\}\geq q$$
$$\Rightarrow \underline{\mathrm{Me}}_M(X)=\inf_{P\in M} \inf\{x:\ P(X\leq x)\geq 0.5\ \wedge\ P(X\geq x)\geq 0.5\}\geq q,$$

but the other direction is not immediately true. So, statements regarding bounds on their imprecise median over a set of distributions are not equivalent to imprecise CDF statements. However, if we start by directly considering,

$$\inf_{P\in M}\ \inf\{x:\ P(X\leq x)\geq 0.5\},$$

instead of $\underline{\mathrm{Me}}_M$, then the statements about the median quantile function over the set of distribution become equivalent to statements about the corresponding set of CDFs. Instead of the formulation by Couso, Moral and Sánchez, we will consider optimising over the *quantile function* as opposed to the end points of the quantile set: this slight modification will result in the upper quantile being defined to maximise (over a set of distributions) the *left* end point of a quantile set instead of the *right*.

Definition 4.3.1: For a set of distributions M and $\alpha\in[0,1]$, we define the *lower and upper imprecise quantile functions of the random variable X over M at level α* as,

$$\underline{\mathrm{Q}}_M^{(\alpha)}(X)=\inf_{P\in M}\ \mathrm{Q}_P^{(\alpha)}(X),$$

$$\overline{\mathrm{Q}}_M^{(\alpha)}(X)=\sup_{P\in M}\ \mathrm{Q}_P^{(\alpha)}(X),$$

with,

$$\mathrm{Q}_P^{(\alpha)}(X)=\inf\{x:\ P(X\leq x)\geq\alpha\}$$

being the quantile function evaluated at the α-th percentile.

Henceforth, the word 'quantile' will refer to the solution of quantile function, which returns the infimum of the set of points where the CDF exceeds α, as oppose to the set itself.

□

Note that the imprecise quantile function in Definition 4.3.1 deviates from the coherence concepts introduced in Chapter 2. Walley's construction of Definition 2.1.8 in [81] is motivated by the geometrical fact that expectations are linear operators on random variables. Because quantiles are not linear operators, it is not readily apparent that this definition is applicable to imprecise quantiles. We do not pursue this avenue any further: in Theorem 4.5.1, we provide a link from the imprecise quantile function in Definition 4.3.1 to a corresponding set of imprecise probabilities.

Unlike the imprecise expectations of Chapter 2, we will be referring to imprecise quantile functions for *unbounded* random variables. To be clear, this is specifically because we will be taking advantage of Theorem 4.5.1 to construct statements about an imprecise quantile function in terms of imprecise probabilities. In turn, imprecise probabilities are expectations of indicator functions of events, which are always bounded random variables. We do not evaluate the imprecise quantile function of unbounded functions in any other fashion.

4.4 Optimisation of imprecise quantile functions

When a convex set of distributions is specified by a generating collection of distributions, it is typically taken to be the closure of its convex hull. In the case of lower and upper expectations, because the expectation operator is a linear functional over the space of distributions, the Krein-Milman theorem (see appendices of Walley [81] or Holmes [47], for example) guarantees that the lower and upper expectations (of *bounded* random variables) are achieved over the generating set. For example, it is typical to optimise over the set of Dirichlet priors when computing imprecise expectations using the IDM (as opposed to finding the optimum over mixtures of the Dirichlet prior).

Because the quantile function is non-linear, we cannot resort to the Krein-Milman theorem. However, because of the monotonicity of the cumulative distribution functions (CDF), we will see that the optimisation of the quantiles over the generating set is the same as the optimisation over its convex hull. The idea of this concept is illustrated as follows.

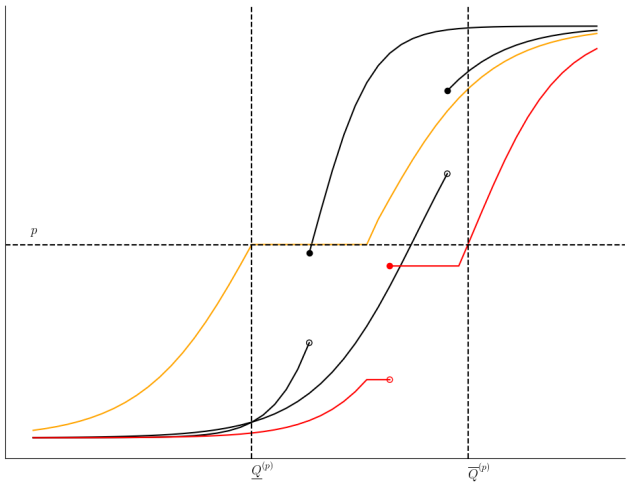

Figure 4.1: Four generating CDF's, along with the lower and upper quantile functions of this set at percentile p. The yellow and red CDF curves respectively represent the minimising and maximising CDF's of the quantile function at p.

In Figure 4.1, a set of four CDF's, M, whose mixture distributions are all the possible CDF's enclosed in the area by the four curves. For a fixed percentile $p \in [0, 1]$ depicted by the horizontal dashed line, the lower quantile function $\underline{Q}^{(p)} = \underline{\mathrm{Q}}^{(p)}_{\mathrm{Conv}(M)}(X)$ of some random variable X is its range value at which any one of the mixtures (and therefore of the four generating CDF's) attain or exceed p. Similarly, the upper quantile function $\overline{Q}^{(p)} = \overline{\mathrm{Q}}^{(p)}_{\mathrm{Conv}(M)}(X)$ is the range value at which all the mixtures CDF's will have attained or exceeded p. The point here is to see that $\underline{Q}^{(p)} = \underline{\mathrm{Q}}^{(p)}_{\mathrm{Conv}(M)}(X)$ is in fact achieved by $\underline{\mathrm{Q}}^{(p)}_{M}(X)$, the minimisation over the generating set instead of the set of mixtures, and similarly for $\overline{Q}^{(p)} = \overline{\mathrm{Q}}^{(p)}_{\mathrm{Conv}(M)}(X)$ by $\overline{\mathrm{Q}}^{(p)}_{M}(X)$. That is, like imprecise expectations, we expect that the optimisation of imprecise quantile functions over a convex set of mixtures is the same as that of over the generating set, which greatly decreases the computational complexity and confirms the intuition provided by the extreme value theorem of convex geometry.

We will now provide a rigourous argument of the above intuition.

Theorem 4.4.1: For a suitably measurable random variable X, let $\{F_a : \ a \in A\}$ be a set of cumulative distribution functions of X indexed by A, a subset of a finite dimensional Euclidean space equipped with the Lebesgue measure. Write the shorthand $\mathrm{Q}_F^{(p)}$ as the p-th quantile function of X under a cumulative distribution function F. Write

$$\Lambda_A = \{\lambda \in \mathbb{R}^A : \ \int \lambda(a)da = 1 \ \wedge \ (\forall a : \lambda(a) \geq 0)\},$$

as a set of mixing functions such that, for a fixed $\lambda \in \Lambda_A$, a mixture over $\{F_a : \ a \in A\}$ is representable as,

$$F_\lambda : \ q \longmapsto \int \lambda(a) F_a(q) da.$$

(We allow λ to be a generalised function such as the Dirac delta function.) Then, for every fixed $p \in [0, 1]$,

$$\underline{\mathrm{Q}}_{\Lambda_A}^{(p)} = \inf_{\lambda \in \Lambda_A} \mathrm{Q}_{F_\lambda}^{(p)} = \inf_{a \in A} \mathrm{Q}_{F_a}^{(p)} = \underline{\mathrm{Q}}_A^{(p)},$$

and

$$\overline{\mathrm{Q}}_{\Lambda_A}^{(p)} = \sup_{\lambda \in \Lambda_A} \mathrm{Q}_{F_\lambda}^{(p)} = \sup_{a \in A} \mathrm{Q}_{F_a}^{(p)} = \overline{\mathrm{Q}}_A^{(p)},$$

Proof: We will prove the infimum case only, as the supremum case can be analogously proven.

Because the minimisation over A is equivalent to minimising over the point mixture masses $\{\delta_a(\cdot) : \ a \in A\}$ which is a subset of Λ_A, the infimum over the latter larger set is smaller than of that over the former smaller subset,

$$\underline{\mathrm{Q}}_{\Lambda_A}^{(p)} = \inf_{\lambda \in \Lambda_A} \mathrm{Q}_{F_\lambda}^{(p)} \leq \inf_{a \in A} \mathrm{Q}_{F_a}^{(p)} = \underline{\mathrm{Q}}_A^{(p)},$$

On the other hand, for any set of distritbutions M and real number q_0, we have,

$$\inf_{F \in M} \mathrm{Q}_F^{(p)} \geq q_0 \iff \forall F \in M : \ \mathrm{Q}_F^{(p)} \geq q_0 \iff \forall F \in M : \ F(q_0) \leq p \iff \sup_{F \in M} F(q_0) \leq p.$$

This is such that, setting $q_0 = \underline{\mathrm{Q}}_A^{(p)}$,

$$\begin{aligned} & \forall a \in A : \ \mathrm{Q}_{F_a}^{(p)} \geq \underline{\mathrm{Q}}_A^{(p)} \\ & \Rightarrow \ \sup_{a \in A} F_a(\underline{\mathrm{Q}}_A^{(p)}) \leq p, \\ & \Rightarrow \ \overline{P}_{M_A}(X \leq \underline{\mathrm{Q}}_A^{(p)}) \leq p. \end{aligned}$$

Recall that an upper probability of an event is an upper expectation of the indicator variable of this event, which in turn is a bounded random variable. Then, the lower envelope theorem 2.2.1 can be applied here, implying that the optimisation over a generated convex set of distributions is the optimisation over the generating set,

$$\begin{aligned}
&\Rightarrow \ \overline{P}_{M_{\Lambda_A}}(X \leq \underline{\mathrm{Q}}_A^{(p)}) \leq p, \\
&\Rightarrow \ \sup_{\lambda \in \Lambda_A} F_\lambda(\underline{\mathrm{Q}}_A^{(p)}) \leq p, \\
&\Rightarrow \ \inf_{\lambda \in \Lambda_A} \mathrm{Q}_{F_\lambda}^{(p)} \geq \underline{\mathrm{Q}}_A^{(p)},
\end{aligned}$$

which is the other side of the inequality, completing the proof.

□

The theorem implies that, for every p for which the quantiles are desired, the minimum and maximum quantiles over the convex set of mixtures over A may be computed by optimising over A directly. Notice that the random variable does *not* need to be bounded in this result: we merely need the fact that imprecise CDF's are imprecise expectations of indicator functions of events *about* X are bounded functions.

4.5 Properties of imprecise quantile functions

Because the quantile function is not necessarily expressible through algebraic operations using expectations, its imprecise version is not necessarily expressible as an optimisation of expectations. Instead, we will rely on the relationship between the quantile and the cumulative distribution function, as follows.

4.5.1 Random variables need not be bounded

First and foremost, we do *not* require that the random variable for which we compute imprecise quantile function values be bounded. Again, this is because the imprecise quantile function is associated with imprecise probabilities: the boundedness requirements of Chapter 2 are automatically satisfied considering that the latter are imprecise expectations of indicator variables that are bounded.

4.5.2 Relation to imprecise probabilities

We now show the connection between the lower and upper imprecise quantile functions over a closed set of distributions M to the lower and upper probabilities over this same set.

Lemma 4.5.1: Let $\underline{P}_M$ and $\overline{P}_M$ be coherent lower and upper probabilities over a closed set of distributions M. Then,

$$\forall P \in M: \; P(A) > \alpha \iff \underline{P}_M(A) > \alpha,$$

$$\forall P \in M: \; P(A) < \alpha \iff \overline{P}_M(A) < \alpha.$$

Proof: We prove the first statement only, since the other can be proven in an analogous manner. It is clear that,

$$\underline{P}_M(A) > \alpha \Rightarrow \forall P \in M: \; P(A) > \alpha.$$

On the other hand,

$$\forall P \in M: \; P(A) > \alpha \iff \underline{P}_M(A) = \inf_{P \in M} P(A) > \alpha.$$

The strictness of the right inequality is given by noting that, from Theorems 2.2.1 and 2.2.2, the infimum on the right is a coherent lower probability and is therefore attainable by an element in M. Strictness follows since the inequality is strict for each element P in M by assumption.

□

Theorem 4.5.1: For $\alpha \in [0,1]$, $q \in \mathbb{R}$, M a closed set of distributions such that $\underline{P}_M$ and $\overline{P}_M$ are coherent, X a suitably measurable random variable,

$$\underline{\mathrm{Q}}_M^{(\alpha)}[X] > q \iff \alpha > \overline{P}_M[X \leq q],$$

$$\underline{\mathrm{Q}}_M^{(\alpha)}[X] \leq q \iff \alpha \leq \overline{P}_M[X \leq q],$$

$$\overline{\mathrm{Q}}_M^{(\alpha)}[X] > q \iff \alpha > \underline{P}_M[X \leq q],$$

and

$$\overline{\mathrm{Q}}_M^{(\alpha)}[X] \leq q \iff \alpha \leq \underline{P}_M[X \leq q].$$

Proof: In order of the statements,

$$
\begin{aligned}
& \underline{\mathrm{Q}}_M^{(\alpha)}[X] > q \\
\iff & \forall P \in M : \ \inf \{x : \ P(X \leq x) \geq \alpha\} > q \\
\iff & \forall P \in M : \ P(X \leq q) < \alpha \\
\iff & \overline{P}_M[X \leq q] < \alpha. && \text{(Lemma 4.5.1)}
\end{aligned}
$$

$$
\begin{aligned}
& \underline{\mathrm{Q}}_M^{(\alpha)}[X] \leq q \\
\iff & \exists P \in M : \ \inf \{x : \ P(X \leq x) \geq \alpha\} \leq q \\
\iff & \exists P \in M : \ P(X \leq q) \geq \alpha \\
\iff & \overline{P}_M[X \leq q] \geq \alpha.
\end{aligned}
$$

$$
\begin{aligned}
& \overline{\mathrm{Q}}_M^{(\alpha)}[X] > q \\
\iff & \exists P \in M : \ \inf \{x : \ P(X \leq x) \geq \alpha\} > q \\
\iff & \exists P \in M : \ P(X \leq q) < \alpha \\
\iff & \underline{P}_M[X \leq q] < \alpha && \text{(Lemma 4.5.1)}
\end{aligned}
$$

$$
\begin{aligned}
& \overline{\mathrm{Q}}_M^{(\alpha)}[X] \leq q \\
\iff & \forall P \in M : \ \inf \{x : \ P(X \leq x) \geq \alpha\} \leq q \\
\iff & \forall P \in M : \ P(X \leq q) \geq \alpha \\
\iff & \underline{P}_M[X \leq q] \geq \alpha
\end{aligned}
$$

□

And so, indeed, inequality statements about the imprecise quantiles are equivalent to lower and upper probabilities of events $\{X \leq x\}$ so there is a direct connection to sets of models

used in statistical applications.

However, they only provide bounds to lower and upper probabilities, thus they are not equivalent to equality statements about coherent values of lower and upper probabilities (recall that a coherent value for, say, a lower probability, is one which achieves the tight minimum over a set of distributions, as opposed to a possibly loose bound).

There are two properties of CDF's that prevent such equality statements. First, CDF's may be flat, such that the quantile function is not a unique inverse function of the CDF. This means that, in general, for any real number q, percentile p, CDF F,

$$F(q) = p \not\Longrightarrow \mathrm{Q}_F^{(p)} = q.$$

Secondly, vertical discontinuities in a CDF may prevent it from achieving all percentiles. That is, evaluating a CDF at its p-th quantile value may not result in p:

$$F(q) = p \not\Longleftarrow \mathrm{Q}_F^{(p)} = q.$$

Conversely, a CDF that does is strictly increasing and continuous everywhere satisfies,

$$F(q) = p \iff \mathrm{Q}_F^{(p)} = q.$$

To generalise this to imprecise probabilities, suppose that, for some fixed q, the minimisation of an imprecise CDF $\underline{F}(q) = \underline{P}(X \leq q)$ is achieved by some F_0 in its set of distributions M at $p = F_0(q)$. That is, by the lower envelope theorem, Theorem 2.2.1, and Theorem 2.2.2 that the lower envelope over M is attainable by an element of M, we have,

$$\forall F \in M: \ F(q) \geq p, \ \exists F_0 \in M: \ F_0(q) = p.$$

If F_0 is also continuous and strictly increasing at q, then,

$$\forall F \in M: \ \mathrm{Q}_F^{(p)} \leq p, \ \exists F_0 \in M: \ \mathrm{Q}_{F_0}^{(p)} = q,$$

is equivalent to the above statement. Finally, Theorem 4.4.1 implies that the above optimisation of the quantile function coincides with the imprecise quantile function over M: that is,

$$\underline{\mathrm{Q}}_M^{(p)} = q.$$

This is the argument for the following result.

Theorem 4.5.2: For $p \in [0,1]$, $q \in \mathbb{R}$, M a closed set of distributions, X a suitably measurable random variable, such that the following optima are achieved by a CDF which is continuous and strictly increasing at q, then,

$$\underline{\mathrm{Q}}_M^{(p)}[X] = q \iff p = \overline{P}_M[X \leq q],$$

and

$$\overline{\mathrm{Q}}_M^{(p)}[X] = q \iff p = \underline{P}_M[X \leq q].$$

□

4.5.3 Lower coverage probability of quantile intervals

We recall that our motivation is to be able to use imprecise quantiles to construct intervals that not only reflect imprecision over a closed set of distributions, but also reflect the variation within each distribution. The following theorem derives a lower bound for the lower coverage probability.

Theorem 4.5.3: For $\alpha, \beta \in [0,1]$ with $\alpha \leq \beta$, $q \in \mathbb{R}$, M a closed set of distributions with continuous CDF's, X a suitably measurable random variable,

$$\underline{P}_M(\underline{\mathrm{Q}}_M^{(\alpha)}[X] \leq X \leq \overline{\mathrm{Q}}_M^{(\beta)}) \geq \beta - \alpha.$$

Proof: Write,

$$\begin{aligned}
&\underline{P}_M(\underline{\mathrm{Q}}_M^{(\alpha)}[X] \leq X \leq \overline{\mathrm{Q}}_M^{(\beta)}) \\
&\quad = \inf\ \{P(\underline{\mathrm{Q}}_M^{(\alpha)} \leq X \leq \overline{\mathrm{Q}}_M^{(\beta)}) :\ P \in M\} \\
&\quad = \inf\ \{P(X \leq \overline{\mathrm{Q}}_M^{(\beta)}) - P(X \leq \underline{\mathrm{Q}}_M^{(\alpha)}) :\ P \in M\}.
\end{aligned}$$

The expression in the infimum is of the form $a_P - b_P$, which is lower bounded by $\inf a_P - \sup b_P$. In other words,

$$\underline{P}_M(\underline{\mathrm{Q}}_M^{(\alpha)}[X] \leq X \leq \overline{\mathrm{Q}}_M^{(\beta)}) \geq \inf\ \{P(X \leq \overline{\mathrm{Q}}_M^{(\beta)}) :\ P \in M\} - \sup\ \{P(X \leq \underline{\mathrm{Q}}_M^{(\alpha)}) :\ P \in M\}.$$

But, by definition,

$$\overline{\mathrm{Q}}_M^{(\beta)}(X) = \sup_{P \in M} \left(\mathrm{Q}_P^{(\beta)}(X) \right).$$

$$\underline{\mathrm{Q}}_M^{(\beta)}(X) = \inf_{P \in M} \mathrm{Q}_P^{(\beta)}(X).$$

Using Theorem 4.5.2 and that the CDF's are continuous by assumption directly yields,

$$\inf_{P \in M} \; P(X \leq \overline{\mathrm{Q}}_M^{(\beta)}) = \beta.$$

$$\sup_{P \in M} \; P(X \leq \underline{\mathrm{Q}}_M^{(\alpha)}) = \alpha.$$

This completes the proof.

□

From an inferential perspective, the coverage probability of the interval $[\underline{\mathrm{Q}}_M^{(\alpha)}, \overline{\mathrm{Q}}_M^{(\beta)}]$ measured with each distributions in M is at least $\beta - \alpha$. In terms of the Bayesian sensitivity analysis methodology, this interval has credibility of at least $\beta - \alpha$ over each distribution in M.

4.6 Hypothesis testing using imprecise quantile intervals

Care should be taken when interpreting inference about questions involving hypotheses of Bayesian probability of zero. A common Bayesian response is that a hypothesis of measure zero cannot be realised exactly, and so is not a valid hypothesis to test when the procedure involves its measure. For example, Gelman et al. [41] and Berger [11] interprets this situation as it is 'unlikely' for a real parameter to be (known to be) precisely an exact real number under continuous distributions. Nevertheless, this hypothesis may represent an important question of study. For example, we may ask: 'Are categorical variables X and Y independent?' Typically, this corresponds to the hypothesis that the their log odds is zero, forming the hypothesis $H = \{\theta : \log \rho(\boldsymbol{\theta}) = 0\}$ for some θ that is multinomial: this hypothesis has measure zero whenever the prior is dominated by the Lebesgue measure.

We will adopt the stance of Gelman et al. [41] and generally consider interpretation of (sets) posterior distributions to be the primary tool of inference. We will report imprecise quantile intervals of a prescribed credibility level and draw qualitative conclusions about the 'point hypothesis' with them. We will also focus on quantile intervals with symmetric percentiles by setting $\alpha = p$ and $\beta = 1 - p$ for $p \in (0, 0.5)$. To wit, suppose $H = \{\boldsymbol{\theta} : \ T(\boldsymbol{\theta}) = t_0\} \subset \Theta$ is of probability zero under a fixed distribution P over Θ. Then, we may compute quantiles satisfying,

$$P(\mathrm{Q}_P^{(p)}(T) \leq T(\theta) \leq \mathrm{Q}_P^{(1-p)}(T)) = 1 - 2p,$$

Then, if $t_0 \in [\mathrm{Q}_P^{(p)}(T), \mathrm{Q}_P^{(1-p)}(T)]$, we interpret this as not having enough evidence to reject $T(\theta) = t_0$ with $1 - 2p$ level credibility. In the imprecise context, we extend this notion by considering whether or not:

$$t_0 \in [\underline{\mathrm{Q}}_M^{(p)}[T], \overline{\mathrm{Q}}_M^{(1-p)}[T]].$$

Due to Theorem 4.5.3, the coverage probability of this interval is at least $1-2p$ over the set of prior distributions M. Thus, if $t_0 \in [\underline{\mathrm{Q}}_M^{(p)}[T], \overline{\mathrm{Q}}_M^{(1-p)}[T]]$, we interpret this as not having enough evidence to reject $T(\theta) = t_0$ with $1 - 2p$ level credibility or higher.

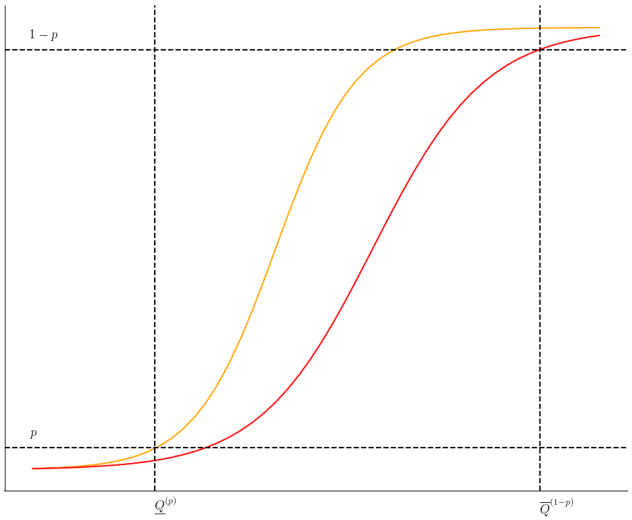

Figure 4.2: Graphical representation of the p-th symmetrical imprecise quantile intervals generated by two distributions and its constituent imprecise quantiles.

From a sensitivity analysis perspective, for $p < 0.5$, the imprecise interval $[\underline{\mathrm{Q}}_M^{(p)}(X), \overline{\mathrm{Q}}_M^{(1-p)}(X)]$ contains all the p-symmetrical intervals $[\mathrm{Q}_P^{(p)}(X), \mathrm{Q}_P^{(1-p)}(X)]$ for all $P \in M$.

4.7 Dataset Examples

We now consider inference arising from using the imprecise quantiles with the IDM. We apply the IDM to the same inferential problem and compare it with the conclusions drawn under the methodology of its original author. To observe the effect of different levels of prior imprecision, we perform inference with increasing values of ν.

Example 4.7.1: Lady Tea Tasting (Fisher [38])

We follow Agresti's [3] of Fisher's [38] classical example of contingency table analysis. This is a small sample table containing the data,

	Guess Milk First	Guess Tea First
Milk First	3	1
Tea First	1	3

For indexing, let ij denote the i-th row and j-th column. The experiment is set up such that the marginal sums are known beforehand: the rows marginals are known because exactly four of cups of teas are poured with milk first and four with tea first, and the columns are known ahead because the lady was told this information before guessing. The parameter of interest is,

$$\rho(\boldsymbol{\theta}) = \frac{\theta_{11}/\theta_{12}}{\theta_{21}/\theta_{22}} = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}}.$$

When all the margins are known, the contingency table follows a hypergeometric sampling. In this case, n_{11} exactly follows a hypergeometric distribution with the probability $P(n_{11} = t)$ interpreted as 'drawing without replacement' from 8 cups of teas with 4 of them having milk first. Furthermore, n_{11} fully determines the table when the margins are given, and, under the null hypothesis of independent sampling $\rho = 1$ such that the hypergeometric distribution is an exact distribution for n_{11}, larger values of the test statistic n_{11} is evidence for $\theta_{11} > 1$ (Agresti [3]).

With this test statistic and null distribution, the one-sided test for,

$$H_0 : \rho = 1, \quad H_a : \rho > 1 \ ,$$

was performed and, for values $n_{11} = 0, 1, 2, 3, 4$ (the number of trials in which the lady guesses correctly), the corresponding p-values are $1.0, 0.986, 0.757, 0.243$ and 0.014 (Agresti [1]): the p-value for this particular table is 0.243 for $n_{11} = 3$.

We now focus on inference on the parameter $\log \rho(\boldsymbol{\theta})$ using the IDM and infer about the null hypothesis $\log \rho(\boldsymbol{\theta}) = 0$ (which is equivalent to $\rho(\boldsymbol{\theta}) = 1$) at a credibility level of 95-%.

	$\nu = 0.5$	1.0	2.0	5.0	10.0
$\underline{Q}^{(0.025)}, \overline{Q}^{(0.025)}$	(-1.457, 0.303)	(-1.82, 0.484)	(-2.232, 0.868)	(-3.288, 1.662)	(-4.274, 2.499)
$\underline{Q}^{(0.975)}, \overline{Q}^{(0.975)}$	(5.329, 8.729)	(4.515, 9.086)	(3.371, 9.212)	(1.549, 9.943)	(0.1, 10.654)

For all tested values of ν, the intervals $[\underline{\mathrm{Q}}_M^{(0.025)}(\log(\rho(\boldsymbol{\theta})), \overline{\mathrm{Q}}_M^{(0.975)}(\log(\rho(\boldsymbol{\theta}))]$ extend further to the right of 0 than to the left of it: this reflects the pattern in the data in which the diagonals are larger than the off-diagonals. In terms of the credibility of H_0, all the intervals $[\underline{\mathrm{Q}}_M^{(0.025)}(\log(\rho(\boldsymbol{\theta})), \overline{\mathrm{Q}}_M^{(0.975)}(\log(\rho(\boldsymbol{\theta}))]$ contain the null value $\theta_0 = 0$.

Interestingly, the imprecise 2.5-% quantiles, $[\underline{\mathrm{Q}}_M^{(0.025)}(\log(\rho(\boldsymbol{\theta})), \overline{\mathrm{Q}}_M^{(0.025)}(\log(\rho(\boldsymbol{\theta}))]$, also all contain the value of $\theta_0 = 0$. That is, some posterior distributions in the IDM are such that $P(\log \rho \leq 0) \leq 0.025$. This suggests that, even at an imprecision level as small as $\nu = 0.5$, there is disagreement amongst the members of the posterior IDM set of distributions in terms of the left side threshold of 2.5-%. In relation to the sensitivity analysis, some priors in the IDM achieve one-sided posterior (Bayesian) p-values of less than 0.025 but some do not. This points to the curious question of which distributions in M yield $\mathrm{Q}_P^{(0.025)}(\log(\rho(\boldsymbol{\theta})))$ that are to the left or to the right of zero: it may be useful to identify similarities amongst prior Dirichlet distributions that induce values on each side of the null hypothesis.

■

In the following examples, we examine what happens when we apply the imprecise quantile functions of the IDM to contingency tables with zero observations in the cells. Due to our

interpreting the imprecise quantile intervals via the coherent imprecise probabilities as in Theorems 4.5.1 and 4.5.2, the validity of computing posterior imprecise probabilities with the IDM over such datasets are dealt with as per the development in Chapter 3, where we detailed the construction of the posterior IDM imprecise expectation for such cases.

Example 4.7.2: Promotion/Racial Bias (Agresti [1], method from Birch [15])

This example illustrates that hypothesis testing with imprecise quantile intervals from the IDM can be highly sensitive to the hyperparameter ν. This is due to the IDM's set of Dirichlet priors consisting of hyperparameter values $\nu\boldsymbol{\alpha}$ over all possible $\boldsymbol{\alpha} \in \triangle^m$, causing it to include prior distributions with $\boldsymbol{\alpha}$ being close to the extreme points of $\triangle^m$. Numerically, we will see that the IDM produces some extreme values for imprecise quantiles for the following log-odds statistic.

Consider the following dataset tabulating the promotion of each person in a group of computer scientists (Gastwirth [40]) based on their race and stratified over three months:

Race	Promotion Month	Not Promoted	Promoted
Black	July	7	0
	August	7	0
	September	8	0
White	July	16	4
	August	13	4
	September	13	2

For indexing, let ijk index over race ($\{B, W\}$), promotion ($\{N, P\}$) and the month ($\{J, A, S\}$), respectively. The goal of this example is to determine the conditional independence of race and promotion variables given the month variable.

Agresti performed the inference under three assumptions. First, the sampling was done independently. (Agresti notes that the information about the overlapping of employee promotion applications over the three months is not available.) Secondly, the test by Birch [15] requires the assumption that the odds ratios over the stratifying variable are equal: that is, for all $k, k' = J, A, S$ in the months variable,

$$\rho_k = \frac{\theta_{BNk}\theta_{WPk}}{\theta_{BPk}\theta_{WNk}} = \frac{\theta_{BNk'}\theta_{WPk'}}{\theta_{BPk'}\theta_{WNk'}} = \rho_{k'}.$$

Finally, Agresti assumes that the marginal counts over race and promotion are fixed. Writing $\rho = \rho_k$ for all k, the hypothesis of interest are

$$H_0 : \rho = 1, \ H_a : \rho < 1.$$

The test of conditional independence by Birch [15] uses the test statistic $n_{BN\cdot}$, whose distribution is a function of the distribution of each n_{BNk}. In turn, the n_{BNk}'s are independent of each other due to the independence assumption and each has the analogous noncentral hypergeometric distribution as in Example 4.7.1. Agresti reports a p-value of 0.026 for this dataset.

Using the IDM, we can compute imprecise quantiles for the log odds,

$$\log \rho(\boldsymbol{\theta}) = \log \frac{\Pr(\text{Black, Not Promoted}) / \Pr(\text{Black, Promoted})}{\Pr(\text{White, Not Promoted}) / \Pr(\text{White, Promoted})} = \log \frac{\theta_{BN\cdot}\theta_{WP\cdot}}{\theta_{BP\cdot}\theta_{WN\cdot}},$$

where, for example,

$$\theta_{11\cdot} = \theta_{11J} + \theta_{11A} + \theta_{11S}.$$

To infer about the point null hypothesis, we consider the lower and upper quantiles at 2.5-% and 97.5-% of $\log \rho(\boldsymbol{\theta})$,

	$\nu = 0.5$	1.0	2.0	5.0	10.0
$\underline{Q}^{(0.025)}, \overline{Q}^{(0.025)}$	$(-\infty$, -7.824)	$(-\infty$, -4.599)	$(-\infty$, -2.652)	$(-\infty$, -1.121)	$(-\infty$, -0.195)
$\underline{Q}^{(0.975)}, \overline{Q}^{(0.975)}$	$(-\infty$, -0.169)	$(-\infty$, 0.232)	$(-\infty$, 0.672)	$(-\infty$, 1.344)	$(-\infty$, 1.896)

The $-\infty$ value taken by all the lower quantiles is consistent with the empty cell $n_{BN\cdot} = 0$. Because $\underline{\mathrm{Q}}_{\mathrm{IDM}}^{(0.975)}$ is at $-\infty$, there exists at least one element of the IDM set of priors whose posterior distribution assigns a probability mass of at least 0.975 to the set $\{\boldsymbol{\theta} : \ \log \rho(\boldsymbol{\theta}) = \log \theta_{BP\cdot} + \log \theta_{WN\cdot} - \log \theta_{BN\cdot} - \log \theta_{WP\cdot} = -\infty\}$.

It can also be seen that between $\nu = 0.5$ and $\nu = 1.0$, the upper quantile changes sign. When $\nu = 0.5$, $\overline{\mathrm{Q}}_{\mathrm{IDM}}^{(0.975)}[\log \rho | \nu, \boldsymbol{n}] = -0.169 < 0$, such that all posterior distributions agree that the value 0 is greater than the 97.5% quantile of $\log \rho$. Together with the lower quantile at 0.025 at $-\infty$, all intervals of the form $[\mathrm{Q}_P^{(0.025)}[\log \rho | \boldsymbol{n}, \nu = 0.5], \mathrm{Q}_P^{(0.975)}[\log \rho | \boldsymbol{n}, \nu = 0.5]]$ over all $P \in M$ are contained in $(-\infty, -0.169]$. This means that all priors produce

$2.5\% - 97.5\%$ quantile intervals that do not contain zero, lending evidence at credibility level 0.05 of rejecting the hypothesis $\log \rho \geq 0$.

On the other hand, for $\nu \geq 1.0$, the imprecise interval $[\underline{\mathrm{Q}}_M^{(0.025)}(\log \rho | \boldsymbol{n}, \nu), \overline{\mathrm{Q}}_M^{(0.975)}(\log \rho | \boldsymbol{n}, \nu)]$ contain 0. Combined with the fact that $\overline{\mathrm{Q}}_M^{(0.025)}(\log \rho | \boldsymbol{n}, \nu) < 0$ for all tested values of ν, it means that there exists at least one $P \in M$ whose $2.5\% - 97.5\%$ quantile interval contains 0. Unlike the case of $\nu = 0.5$, there is no definite consensus amongst the priors for $\nu \geq 1.0$. In other words, at $\nu = 0.5$, all priors lend credibility to the system being not independent at level 97.5%, roughly in agreement with the classical test. However, when $\nu \geq 1.0$, at least one prior will not yield this conclusion. This demonstrates the sensitivity of the inference to the IDM hyperparameter ν, as discussed at the beginning of this example.

■

Example 4.7.3: (Probability of a girl birth given placenta previa [41])

This example illustrates the behaviour of the IDM imprecise quantiles when the sample size is large and the number of observations in all categories are far from zero. It also compares the inference from an IDM model to the inference from a Bayesian treatment using a single Dirichlet prior. Because of this, the example also highlights the fact that the IDM inference includes a global analysis over the family of models considered, as opposed to the local analysis performed on a finite number of models *post-hoc* of Bayesian inference.

Placenta previa is a 'condition of pregnancy in which the placenta is ... obstructing the fetus from normal vaginal delivery.' Gelman et al. [41] cite a data set of placenta previa births from Germany with 437 female births out of 980, and they are interested in the inference of the proportion of female births under this condition, and the odds ratio between the births two genders under this condition. In particular, they wish to test the hypothesis that the proportion is less than 0.485.

One of their precise Bayesian treatments is as follows. Let θ be the probability of a female birth under the placenta previa condition. A $\mathrm{Beta}(1, 1)$ prior distribution was used with a binomial likelihood: using the Germany dataset, the posterior distribution for θ is $\mathrm{Beta}(438, 544)$. The statistics of interest are θ and $(1 - \theta)/\theta$. The authors report 95% posterior intervals of $[0.415, 0.477]$ and $[1.10, 1.41]$ for θ and $(1 - \theta)/\theta$, respectively.

Sensitivity analysis is done by picking models which are increasingly concentrated around 0.485, the value of the hypothesis of interest: Gelman et al. fixed the prior expectation $\alpha/(\alpha+\beta)$ to 0.485 and chose the value of $\alpha+\beta$ to be $2, 5, 10, 20, 100, 200$. They report that all cases except for the last two are models whose '[p]osterior inferences based on a large sample are not particularly sensitive to the prior distribution' [41]. In the last two cases, where the prior information is strong, they notice the 'posterior intervals (being) pulled noticeably toward the prior distribution' [41] but that still, 'the 95% posterior intervals still exclude the prior mean' [41]. Their sensitivity analyses is summarised by the following table from the author:

$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	95% (central) posterior interval for θ
0.500	2	[0.415, 0.477]
0.485	2	[0.415, 0.477]
0.485	5	[0.415, 0.477]
0.485	10	[0.415, 0.477]
0.485	20	[0.416, 0.478]
0.485	100	[0.420, 0.479]

Figure 4.3: Sensitivity analysis from the example of Gelman et al. [41].

We compare this example with the IDM of two classes (also known as the *Imprecise Beta model (IBM)*). The imprecise expectation and quantiles of $\theta = \Pr(\text{Birth under Placenta Previa is girl})$, using values $\nu = 2, 5, 10, 20, 100$ are:

	$\nu = 2.0$	5.0	10.0	20.0	100.0
$\underline{E}(\theta), \overline{E}(\theta)$	(0.445, 0.447)	(0.444, 0.449)	(0.441, 0.452)	(0.437, 0.457)	(0.405, 0.497)
$\underline{Q}^{(0.025)}(\theta), \overline{Q}^{(0.025)}(\theta)$	(0.408, 0.422)	(0.406, 0.423)	(0.404, 0.427)	(0.399, 0.432)	(0.369, 0.473)
$\underline{Q}^{(0.05)}(\theta), \overline{Q}^{(0.05)}(\theta)$	(0.414, 0.426)	(0.413, 0.428)	(0.411, 0.431)	(0.406, 0.436)	(0.375, 0.477)
$\underline{Q}^{(0.5)}(\theta), \overline{Q}^{(0.5)}(\theta)$	(0.442, 0.45)	(0.441, 0.451)	(0.438, 0.454)	(0.434, 0.46)	(0.402, 0.5)
$\underline{Q}^{(0.95)}(\theta), \overline{Q}^{(0.95)}(\theta)$	(0.466, 0.478)	(0.465, 0.48)	(0.463, 0.483)	(0.458, 0.488)	(0.425, 0.528)
$\underline{Q}^{(0.975)}(\theta), \overline{Q}^{(0.975)}(\theta)$	(0.47, 0.485)	(0.469, 0.487)	(0.466, 0.489)	(0.462, 0.494)	(0.428, 0.533)

We compare the 2.5-th, 50-th, and 97.5-th imprecise quantiles to the 95-th central interval (centred at the 50-th percentile) reported by Gelman et. al. We recall that ν is the sum of the Beta parameters for all the prior distributions in the set of distributions.

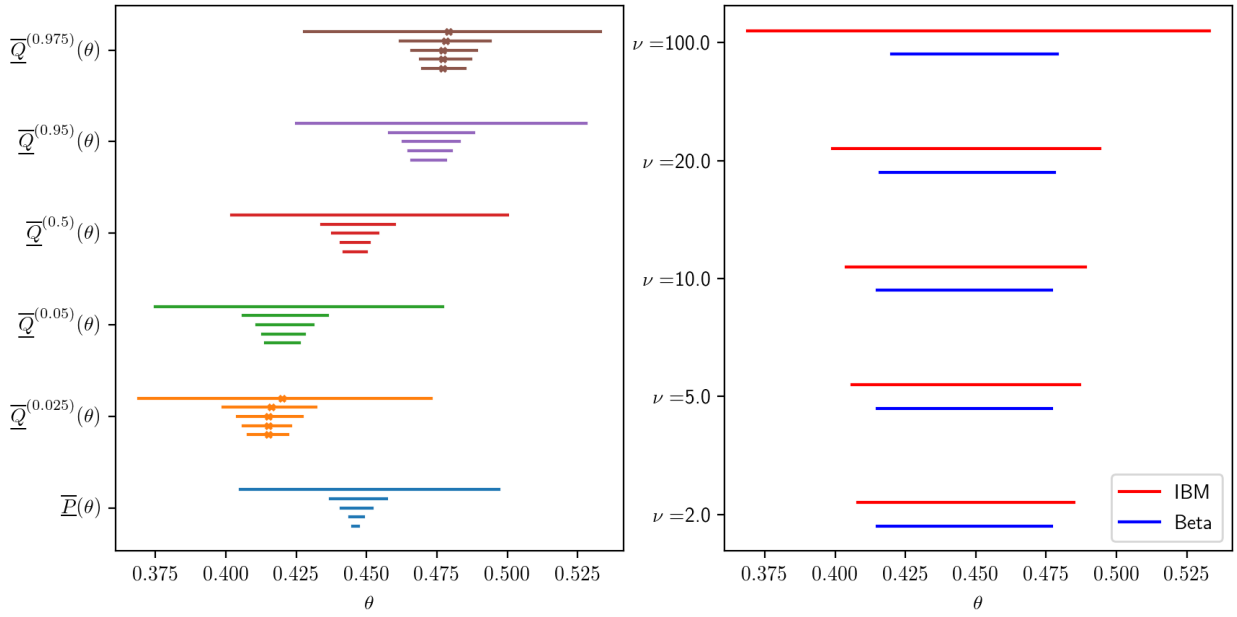

Figure 4.4: Left: Imprecise intervals, $\overline{Q}^{(\alpha)}(\theta) := [\underline{Q}^{(\alpha)}_{\text{IDM}}(\theta|\nu,\boldsymbol{n}), \overline{Q}^{(\alpha)}_{\text{IDM}}(\theta|\nu,\boldsymbol{n})]$ for various values of ν. The imprecise expectations $\overline{E}(\theta) := [\underline{E}_{\text{IDM}}(\theta|\nu,\boldsymbol{n}), \overline{E}_{\text{IDM}}(\theta|\nu,\boldsymbol{n})]$ are also plotted. (Shorter lengths indicate lower ν values in 2,5,10,20,100). The left and right bounds of the precise Beta interval from Gelman et al. [41] are marked for the 2.5 and 97.5 percentiles. Right: Plot of the imprecise interval $[\underline{\text{Q}}^{(0.025)}_{\text{IDM}}, \overline{\text{Q}}^{(0.975)}_{\text{IDM}}]$ and the precise Beta intervals $[\text{Q}^{(0.025)}_{\text{Beta}}, \text{Q}^{(0.975)}_{\text{Beta}}]$ from Gelman et al. for different ν values.

From the left plot of Figure 4.4, an interesting property regarding an imprecise model such as the IDM is that the intervals representing imprecise quantiles of differing percentiles can overlap when imprecision is large: for example, at $\nu = 20$, $\overline{\text{Q}}^{(0.05)}(\theta) \approx 0.436 > 0.434 \approx \underline{\text{Q}}^{(0.5)}(\theta)$, meaning that there exists two posterior Beta distributions in the set of distributions such that the 5-th quantile of one is greater than 50-th quantile of the other, with the maximiser of the 5-th quantile placing significantly more mass to the right of 0.436 than the minimiser of the 50-th quantile. In addition, we similarly observe the overlap $\overline{\text{Q}}^{(0.5)} \approx 0.46 > 0.458 \approx \underline{\text{Q}}^{(0.975)}$. These suggest that the posterior set of distributions contains Beta distributions where $\alpha > \beta$ and $\alpha < \beta$, indicating that there is no agreement to the (a)symmetry of the shape of the Beta posteriors. This means that the data, despite containing many samples of a simple Bernoulli system, fails to distinguish between these two types of Beta distributions since they are both still consistent with the data.

■

Example 4.7.4: Opinion on government spending on the arts (Agresti [2], Table 11.1, Section 11.2.5)

This example again compares the IDM to the Bayesian treatment with a single Dirchlet prior. We examine the effect of imprecision on inference when the dataset is both small-sample, contains a zero observation one category and one observation in another. We will see that IDM yields a range of posterior values which are consistent with the model's set of Dirichlet priors: these values will turn out to be quite different than the values obtained by Agresti using a single prior. This reflects the sensitivity to prior specification for inference in situations such as this.

The following dataset describes the opinions of 23 female survey takers of ages 18 to 21 on government spending on culture and the arts.

Opinion	Much More	More	Same	Less	Much Less
	1	7	12	3	0

Note that not only is this a somewhat small sample, but also the data contains a cell with zero observations (the 'Much Less' category).

Agresti models the multinomial cell probabilities with a Dirichlet prior with equal relative weights differing concentrations: the goal is to regularise the cell probabilities estimates in light of the small sample size and having zero observation in some categories. Note that the prior implies that the modeller believes that, in expectation, all cell probabilities are equal. The Bayes' point estimates are given as follows (Agresti [2]),

	Much More	More	Same	Less	Much Less
$\nu = 1$	0.050	0.300	0.508	0.133	0.008
$\nu = 5$	0.071	0.286	0.464	0.143	0.036
$\nu = 20$	0.116	0.256	0.372	0.163	0.093

(Credible intervals were not provided by Agresti.) We run the IDM inference also using $\nu = 1, 5, 20$ for the imprecise expectations of the cell probabilities:

Opinion	Much More	More	Same	Less	Much Less
$\underline{E}(\theta), \overline{E}(\theta)$	(0.042, 0.083)	(0.292, 0.333)	(0.5, 0.542)	(0.125, 0.167)	(0.0, 0.042)
$\underline{Q}^{(0.025)}(\theta), \overline{Q}^{(0.025)}(\theta)$	(0.0, 0.016)	(0.109, 0.191)	(0.273, 0.378)	(0.019, 0.064)	(0.0, 0.002)
$\underline{Q}^{(0.975)}(\theta), \overline{Q}^{(0.975)}(\theta)$	(0.12, 0.258)	(0.445, 0.57)	(0.658, 0.766)	(0.248, 0.376)	(0.0, 0.18)

Opinion	Much More	More	Same	Less	Much Less
$\underline{E}(\theta), \overline{E}(\theta)$	(0.036, 0.214)	(0.25, 0.429)	(0.427, 0.607)	(0.107, 0.285)	(0.0, 0.178)
$\underline{Q}^{(0.025)}(\theta), \overline{Q}^{(0.025)}(\theta)$	(0.0, 0.102)	(0.093, 0.281)	(0.225, 0.454)	(0.016, 0.16)	(0.0, 0.077)
$\underline{Q}^{(0.975)}(\theta), \overline{Q}^{(0.975)}(\theta)$	(0.103, 0.424)	(0.388, 0.647)	(0.576, 0.802)	(0.214, 0.499)	(0.0, 0.374)

Opinion	Much More	More	Same	Less	Much Less
$\underline{E}(\theta), \overline{E}(\theta)$	(0.023, 0.488)	(0.161, 0.628)	(0.275, 0.744)	(0.07, 0.535)	(0.0, 0.465)
$\underline{Q}^{(0.025)}(\theta), \overline{Q}^{(0.025)}(\theta)$	(0.0, 0.365)	(0.058, 0.508)	(0.136, 0.633)	(0.011, 0.412)	(0.0, 0.345)
$\underline{Q}^{(0.975)}(\theta), \overline{Q}^{(0.975)}(\theta)$	(0.066, 0.663)	(0.258, 0.787)	(0.386, 0.879)	(0.14, 0.704)	(0.0, 0.641)

As a sanity check, note that all the posterior expectations from Agresti's calculations are strictly in the imprecise expectations of the IDM, as expected. We now focus on the inference provided by the 2.5% imprecise quantiles of the IDM for the cells 'Much More' and 'Much Less', the former with a single observation and the latter with no observations.

The lower expectation of 'Much Less' is zero which is in accordance with the formula,

$$\underline{E}_{\text{IDM}}(\theta_i|\boldsymbol{n}, \nu) = \inf \left\{ \frac{\nu\alpha_i + n_i}{\nu + n} : \alpha \in \overline{\triangle^p} \right\} = \frac{n_i}{\nu + n},$$

with $n_i = 0$. Since we are considering the extended parameter space, this lower bound is achievable by any Dirichlet distribution such that $\alpha_i = 0$. For such a Dirichlet distribution, the cell probability θ_i is almost surely zero since the expectation of a nonnegative bounded random variable is zero. This forces all the quantiles of this distribution to equal zero, and therefore the distribution also achieves the minimum of all imprecise quantiles of this cell probability. This corroborates the fact that the lower quantiles at both 2.5% and 97.5% are zero at all values of ν for 'Much Less'.

On the other hand, the single observation in 'Much More' causes the lower expectation of its cell probability to be non-zero, again by the lower expectation formula. Note that for this category, Agresti's point estimates always tend to be closer to $\underline{E}(\theta_{\text{Much More}}|\nu, \boldsymbol{n})$ than

to $\overline{E}(\theta_{\text{Much More}}|\nu, \boldsymbol{n})$, the former of which being brought away from zero due to the single observation. This corroborates the possibility that, since Agresti started with a uniform Dirichlet distribution, the low cell count in the data relative to other categories is reflected here. However, this also means that the fact that there are other posterior expectations up to $\overline{E}(\theta)$ is also consistent with the global sensitivity analysis it provides. The inference is therefore sensitive to the prior as there is only one count in the cell of interest, and this is reflected by the imprecise expectation $[\underline{E}_{\text{IDM}}(\theta_{\text{Much More}}|\nu, \boldsymbol{n}), \overline{E}_{\text{IDM}}(\theta_{\text{Much More}}|\nu, \boldsymbol{n})]$ containing values which are very different from the single point estimate of using a single uniform Dirichlet prior as in Agresti.

■

4.8 Concluding remarks

Our contributions in this chapters are as follows. We define the *imprecise quantile function*, the centerpiece of this chapter, in Definition 4.3.1. Theorems 4.4.1, 4.5.1 and 4.5.2 describe some of its imprecise-probabilistic properties. Theorem 4.5.3 guarantees a lower bound of the coverage probability of an interval using quantiles derived from the imprecise quantile function. Section 4.6 discusses the interpretation of the imprecise quantile function in prior/posterior inference settings, while Examples 4.7.1, 4.7.2, 4.7.3, 4.7.4 demonstrate its application on real-life datasets.

We have defined the imprecise analogue of a quantile function and shown that, like the imprecise expectations, the optimisation of the lower and upper quantile functions over a set of mixtures induced by some set of distributions M can be achieved within M. Because the quantile function is a nonlinear functional and, in a dual manner, CDF's of sums do not generally decompose into sums of CDF's of the component random variables, the imprecise quantile is not necessarily coherent. Nevertheless, we noted that they are valuable for interpretation in two ways: they are properly a sensitivity analysis of the quantile function over a set of distributions and that they have logically equivalent statements in terms of bounds on the lower and upper CDF's. In the case that the CDF's are all continuous, these bounds are tight and statements about imprecise quantiles are exactly statements about lower and upper CDF's, not just their bounds (Theorem 4.5.2).

Further, we used such imprecise quantiles to construct imprecise *interval* statistics that has

the sensitivity analysis interpretation of containing all the quantile intervals induced by the mixture of the generating set of distributions. Unlike the imprecise expectation, which is a set of point estimates over the inducing set of distributions, imprecise interval formed using quantiles also measures variations due to the *randomness* contributed by each distribution as well. Furthermore, despite the quantiles themselves not necessarily coherent, we have shown that the lower coverage probability of our imprecise quantile interval using the α-th and β-th quantiles ($\alpha < \beta$) is at least lower bounded by the nominal coverage probability $\beta - \alpha$ of each model in the set of distributions. This provides an imprecise analogue to the use of credible intervals in the precise Bayesian framework. We have compared our methodology to existing methodologies of other authors.

It is interesting to note that the construction of the imprecise quantile functions is an example of coherence departing from sensitivity analysis. By this, we mean that, unlike the imprecise expectation where the lower envelope theorem (Theorem 2.2.1) implies that coherence is equivalent to tight bounds of sensitivity analysis (that is, minimisation and maximisation) of expectations over a set of distributions, our construction of imprecise quantiles demonstrates that this need not hold, at least for sensitivity analysis of nonlinear functionals.

This represents a divergence between the adherence to coherence in Bayesian analysis (in the imprecise setting) and sensitivity analysis. In practice, one would like to do the latter, and the imprecise framework allows one to perform such analysis globally on a set of candidate (prior) distributions (and such analysis is not post-hoc as it is part of the inferential process in the imprecise framework). On the other hand, in principle, the imprecise quantile functions themselves cannot be used as a coherent bounds on from a behavioural perspective. (See Couso, Moral and Sánchez [30] and Couso and Dubois [28] that generalise the notion of *gambling* and *desirability* beyond bounds on expectations as coherent assessments.)

Chapter 5

On the optimisation problem for the log-odds inference with IDM

An integral part of inference using imprecise expectations and probabilities is the optimisation over the elicited set of priors. and it becomes important to characterise the set of optima as part of the qualitative understanding of one's inference. In this chapter, we study the specific optimisation problem for the lower expectation of log-odds statistics under the IDM model. Under the Dirichlet assumption, the expectation of log-odds is a linear combination of polygamma functions which are individually continuously differentiable and monotonic (see Appendix D.2). However, the difference of such functions is more involved than expected from working with such nice functions individually. We will highlight key properties of two aspects of the problem: the identification of a subset of the natural parameters using the Karush-Kuhn-Tucker (KKT) conditions and the properties of the objective function over this subset.

5.1 KKT solutions to common log-odds problems

In this section, we study the Karush-Kuhn-Tucker (KKT) conditions of three classes of log-odds that commonly appear in statistical applications. In particular, we study how knowing the specific structure of the log-odds may further reduce the set of candidate solutions to the optimisation problem.

5.1.1 An overview of the KKT conditions

The KKT conditions (for example, see [18]) characterise the necessary conditions of a constrained optimisation problem,

$$\begin{aligned} \text{minimise: } & I(\boldsymbol{\alpha}), \\ \text{subjected to: } & \boldsymbol{g}(\boldsymbol{\alpha}) \leq \mathbf{0}, \\ & h(\boldsymbol{\alpha}) = 0, \end{aligned}$$

where $I : \mathbb{R}^m \longmapsto \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^l$ and $h : \mathbb{R}^m \to \mathbb{R}$ represent the objective function, l inequality constraints and a single equality constraint respectively. For the inequality and equality constraints, a vector $\boldsymbol{\mu} \in \mathbb{R}^l$ and scalar $\lambda \in \mathbb{R}$ of multipliers are introduced respectively. The KKT necessary conditions for $(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*, \lambda^*)$ to be a constrained minimiser of this problem are,

$$\begin{gathered} \nabla I(\boldsymbol{\alpha}^*) + \sum_{k=1}^{l} g_k \nabla \boldsymbol{g}(\boldsymbol{\alpha}^*) + \lambda^* \nabla h(\boldsymbol{\alpha}^*) = 0 \\ \forall i \in 1, \ldots, l : \ \mu_i^* g_i(\boldsymbol{\alpha}^*) = 0, \\ \boldsymbol{g}(\boldsymbol{\alpha}^*) \leq 0, \\ h(\boldsymbol{\alpha}^*) = 0, \\ \boldsymbol{\mu}^* \geq \mathbf{0}. \end{gathered}$$

In particular, the constraints that $\boldsymbol{\alpha}^*$ is in the m-simplex is written as,

$$\begin{gathered} \forall i \in 1, \ldots, m : \ \mu_i^* \alpha_i^* = 0, \\ \forall i \in 1, \ldots, m : \ \alpha_i^* \geq 0, \\ 1 - \sum_{i=1}^{m} \alpha_i^* = 0, \\ \boldsymbol{\mu}^* \geq \mathbf{0}. \end{gathered}$$

The conditions $\mu_i^* \alpha_i^* = 0$ will be used extensively in what follows: a strategy for reducing the solution space using the KKT condition is to identify $\alpha_i^* = 0$ by showing that $\mu_i^* > 0$ using the properties of the stationariry conditions $I(\boldsymbol{\alpha}^*) + \sum_{k=1}^{l} g_k \nabla \boldsymbol{g}(\boldsymbol{\alpha}^*) + \lambda^* \nabla h(\boldsymbol{\alpha}^*) = 0$.

5.1.2 The KKT conditions of posterior log-odds lower expectation under the IDM

We will use two main facts to drive our analyses: the trigamma function ψ' is strictly positive over the positive real numbers (see Appendix D.2) and the optimal values of the complementary KKT multipliers $\boldsymbol{\mu}$ are non-negative (see, for example, Boyd and Vandenberghe [18]).

Let L be a finite number of categories of the observation multinomial model. The general form of the IDM posterior lower expectation of the log-odds ratio between two finite collections $\mathcal{A}$ and $\mathcal{B}$ of subsets of L,

$$\begin{aligned}&\underline{E}_{\text{IDM}}\left(\sum_{A\in\mathcal{A}}\log\theta_A-\sum_{B\in\mathcal{B}}\log\theta_B\;\middle|\;\nu,\boldsymbol{n}\right)\\&\quad=\min\left\{\sum_{A\in\mathcal{A}}\psi(\nu\alpha_A+n_A)-\sum_{B\in\mathcal{B}}\psi(\nu\alpha_B+n_B):\;\boldsymbol{\alpha}\in\triangle^L\right\},\end{aligned}$$

where ψ is the digamma function and, for a vector $\boldsymbol{a}=(a_1,\ldots,a_m)$ and a subset $C\subseteq\{1,\ldots,m\}$,

$$a_C:=\sum_{i\in C}a_i.$$

We focus on cases where $|\mathcal{A}|,|\mathcal{B}|$ are at most two, so that we are interested in minimising the function,

$$\boldsymbol{\alpha}\longmapsto\;\psi(\nu\alpha_{A_1}+n_{A_1})+\psi(\nu\alpha_{A_2}+n_{A_2})-\psi(\nu\alpha_{B_1}+n_{B_1})-\psi(\nu\alpha_{B_2}+n_{B_2}).$$

The Lagrangian equations of the KKT conditions for the minimisers of this problem are,

$$\begin{gathered}\nabla_{\boldsymbol{\alpha}^*}\left(\psi'(\nu\alpha^*_{A_1}+n_{A_1})+\psi(\nu\alpha^*_{A_2}+n_{A_2})\right.\\\left.-\psi(\nu\alpha^*_{B_1}+n_{B_1})-\psi(\nu\alpha^*_{B_2}+n_{B_2})\right)\\-\boldsymbol{\mu}^*+\lambda^*\mathbf{1}=0.\end{gathered}$$

and ψ' is the trigamma function. We will sometimes denote the digamma and trigamma functions evaluated at a KKT solution as,

$$\psi_C:=\psi(\nu\alpha^*_C+n_C),$$

and

$$\psi'_C := \psi'(\nu\alpha^*_C + n_C),$$

respectively, when we do not need to appeal to the arguments within and wish to simplify notation.

Notice that, because,

$$\frac{\partial}{\partial\alpha_i}\alpha_C,$$

is zero when $i \notin C$, some trigamma function terms of the Lagrangian equations in some α_i's will become zero depending on whether or not that α_i lies in A_1, A_2, B_1, B_2 or not. In particular, it is useful to distinguish the cases when a category k falls into the numerator or denominator sets:

$$\begin{aligned}
k \in A_1 \cup A_2,\ k \in B_1 \cup B_2 &: \psi'_{A_1}[k \in A_1] + \psi'_{A_2}[k \in A_2] - \psi'_{B_1}[k \in B_1] - \psi'_{B_2}[k \in B_2] - \mu_k + \lambda = 0,\\
k' \in A_1 \cup A_2,\ k' \notin B_1 \cup B_2 &: \psi'_{A_1}[k' \in A_1] + \psi'_{A_2}[k' \in A_2] - \mu'_k + \lambda = 0,\\
k'' \notin A_1 \cup A_2,\ k'' \in B_1 \cup B_2 &: -\psi'_{B_1}[k'' \in B_1] - \psi'_{B_2}[k'' \in B_2] - \mu''_k + \lambda = 0,\\
k''' \notin A_1 \cup A_2,\ k''' \notin B_1 \cup B_2 &: -\mu'''_k + \lambda = 0,
\end{aligned}$$

where $[k \in C]$ is the indicator function that k is in C. Depending on A_1, A_2, B_1, B_2, L, some of these forms may not be present. For example, if the log-odds involves all the categories in question, such that $A_1 \cup A_2 \cup B_1 \cup B_2 = L$, then the Lagrangian equation of the form k''' will not be present. Similarly, if $A_1 \cup A_2 \subset B_1 \cup B_2$, then equations of the class of k' will not be present.

The significance of this decomposition is that, for some equations, the positivity of the trigamma function that is uniform over $\triangle^L \ni \boldsymbol{\alpha}$ can be more readily exploited when they are present. In particular, we will frequently make use of the following results.

Lemma 5.1.1: When there exists $k' \in L$ such that $k' \in A_1 \cup A_2$ and $k' \notin B_1 \cup B_2$,

$$\lambda < \mu_{k'}.$$

Proof: This follows from,

$$0 < \psi'_{A_1}[k \in A_1] + \psi'_{A_2}[k \in A_2] = \mu_{k'} - \lambda.$$

□

Lemma 5.1.2: When there exists $k'' \in L$ such that $k'' \notin A_1 \cup A_2$ and $k'' \in B_1 \cup B_2$,

$$\lambda > 0.$$

Proof: This follows from,

$$0 < \psi'_{B_1}[k \in B_1] + \psi'_{B_2}[k \in B_2] + \mu_{k''} < \lambda.$$

□

Lemma 5.1.3: When there exists $k''' \in L$ such that $k''' \notin A_1 \cup A_2$ and $k''' \notin B_1 \cup B_2$,

$$\mu_{k'''} = \lambda.$$

Consequently,

$$\lambda \geq 0.$$

□

Then, for example, if k', k'', k''' exist, then we can conclude that $\mu_{k'}, \mu_{k'''} > 0$, leading to the minimiser satisfying $\alpha^*_{k'}, \alpha^*_{k'''} = 0$ by the KKT complementary conditions.

As we have alluded at the beginning, we want to use the Lagrangian conditions,

$$\begin{aligned}\boldsymbol{\mu}^* = \lambda^* \mathbf{1} + \nabla_{\boldsymbol{\alpha}^*} \big(&\psi'(\nu\alpha^*_{A_1} + n_{A_1}) + \psi(\nu\alpha^*_{A_2} + n_{A_2}) \\ &-\psi(\nu\alpha^*_{B_1} + n_{B_1}) - \psi(\nu\alpha^*_{B_2} + n_{B_2}) \big).\end{aligned}$$

by identifying which components μ^*_k are strictly greater than zero to conclude that, correspondingly using the slackness condition $\mu^*_k \alpha^*_k = 0$, $\alpha^*_k = 0$. To that end, define,

$$d_k(\boldsymbol{\alpha}; \boldsymbol{n}, \nu) := \psi'_{A_1}(\boldsymbol{\alpha}; \boldsymbol{n}, \nu)[k \in A_1] + \psi'_{A_2}(\boldsymbol{\alpha}; \boldsymbol{n}, \nu)[k \in A_2] - \psi'_{B_1}(\boldsymbol{\alpha}; \boldsymbol{n}, \nu)[k \in B_1] - \psi'_{B_2}, (\boldsymbol{\alpha}; \boldsymbol{n}, \nu)[k \in B_2].$$

(We will write $d_k(\boldsymbol{\alpha})$ when $\boldsymbol{n}$ and ν are notationally unambiguous.) This is such that the k-th Lagrangian equation can be written succinctly as,

$$\mu^*_k = \lambda^* + d_k(\boldsymbol{\alpha}^*; \boldsymbol{n}, \nu).$$

An equivalent sufficient condition for $\alpha^*_k = 0$ is that $d_k(\boldsymbol{\alpha}^*; \boldsymbol{n}, \nu) + \lambda^*$. This is useful when, say $\lambda^* \geq 0$ such that it is sufficient to show $d_k > 0$ to conclude $\mu^*_k > 0$. In these cases, the

properties of the function $d_k(\boldsymbol{\alpha}|\boldsymbol{n}, \nu)$ over $\boldsymbol{\alpha}$ parametrised by the data $\boldsymbol{n}$ and the hyperparameter ν becomes of great interest to us.

Note that the sign of $d_k(\boldsymbol{\alpha}^*)$, a difference of sums of trigamma functions, is generally dependent on $\boldsymbol{\alpha}$ and $\boldsymbol{n}$. For a fixed set of parameters $\boldsymbol{n}$, ν and a particular k in question, suppose $I^* = \{i \in 1, \ldots, L : \ \alpha_i^* = 0\}$ has been identified by, for example, the KKT conditions. Then, our solution space is reduced to the face,

$$\{\boldsymbol{\alpha} \in \triangle^L : \ \forall i \in I^*, \ \alpha_i = 0.\}.$$

Now if $d_k > 0$ over this face, then the Lagrangian equation of k can be exploited in a similar manner to the others: setting $\alpha_i = 0$ for $i \in I^*$, we can write, for any $\boldsymbol{\alpha}^*$ that is a KKT solution in $\{\boldsymbol{\alpha} \in \triangle^L : \ \forall i \in I^*, \ \alpha_i = 0.\}$,

$$0 < d_k(\boldsymbol{\alpha}^*) = \mu_k^* - \lambda^*.$$

leading to,

$$\lambda^* < \mu_k^*.$$

Otherwise, the KKT complementary slackness condition is not used, and we resort to other methods of analysis depending on the specifics of the problem. For example, we may identify additional solutions to the constrained optimisation by noticing that d_k is in fact the k-th element of the gradient of the objective function (scaled by $1/\nu$) and considering the properties of this function over the face identified by I^*.

5.1.3 Log probability ratios

To illustrate the general approach, we look at the most important examples of odds ratios statistics.

Let L be a finite number of categories, $A, B \subseteq L$, $\nu > 0$ be the predefined hyperparameter of the IDM, $\boldsymbol{\theta} \in \triangle^L$ be a multinomial probability vector over L, and $\boldsymbol{n} \in \mathbb{N}^L$ be a vector of non-zero counts. We are interested in computing,

$$\underline{E}_{\text{IDM}}\left(\log \frac{\theta_A}{\theta_B} \,\middle|\, \nu, \boldsymbol{n}\right) = \psi(\nu\alpha_A + n_A) - \psi(\nu\alpha_B + n_B),$$

where $\theta_A = \sum_{l \in A} \theta_l$ is the sampling probability of the set A.

The possible KKT stationarity conditions for this problem are,

$$\begin{aligned}
\forall k \in A,\ k \in B : \nu\psi'_A - \nu\psi'_B - \mu_k + \lambda = 0 \\
\forall k' \in A,\ k' \notin B : \nu\psi'_A - \mu_{k'} + \lambda = 0, \\
\forall k'' \notin A,\ k'' \in B : -\nu\psi'_B - \mu_{k''} + \lambda = 0, \\
\forall k''' \notin A,\ k''' \notin B : -\mu_{k'''} + \lambda = 0.
\end{aligned}$$

From Lemmas 5.1.1, 5.1.2 and 5.1.3, when the corresponding equations exist, we can deduce the following algebraic consequences of each of the equations.

$$\begin{aligned}
\forall k' \in A,\ k' \notin B : \ \mu_{k'} > \lambda. \\
\forall k'' \notin A,\ k'' \in B : \ \lambda > 0. \\
\forall k''' \notin A,\ k''' \notin B : \ \mu_{k'''} = \lambda.
\end{aligned}$$

More information is needed to analyse the μ's of two classes k and k'', and we will do so by a specific problem.

Example 5.1.1: Suppose $L = \{1, 2, 3, 4\}$, $A = \{1, 2\}$ and $B = \{2, 3\}$ and fix a hyperparameter ν and dataset $\boldsymbol{n}$. Now,

$$k \in A \cap B = \{2\},\ k' \in A \cap B^c = \{1\},\ k'' \in A^c \cap B = \{3\},\ k''' \in (A \cup B)^c = \{1, 2, 3\}^c = \{4\},$$

From $k'' = 3$, we deduce,

$$\lambda > 0.$$

This is such that $\mu_1, \mu_4 > 0$ and so,

$$\alpha_1^*, \alpha_4^* = 0.$$

So, our solution space is now restricted to the face consisting of points $(0, \alpha_2, \alpha_3, 0) \in \triangle^L$, such that $\alpha_3^* = 1 - \alpha_2^*$.

In this particularly simple problem, we can directly appeal to the objective function in this reduced parametrisation,

$$\alpha_2 \longmapsto \psi(\nu\alpha_2 + n_1 + n_2) - \psi(\nu + n_2 + n_3),$$

and, by the increasing property of the digamma function ψ, the minimum must be $\alpha_2^* = 0$, leading to $\alpha_3^* = 1$.

■

Let us also point out a general pattern in the log-probability problem.

Theorem 5.1.1: Consider non-empty sets $A \neq B \subset L$ such that the sets $A \cap B^c$ and $A^c \cap B$ are nonempty. Then, all minimisers $\boldsymbol{\alpha}^*$ of the log-probability IDM problem lie on the face of the simplex $\triangle^L$ satisfying

$$\alpha_{A^c \cap B} = 1.$$

Proof: Given non-empty sets $A \neq B \subset L$ such that the sets $A \cap B^c$ and $A^c \cap B$ are nonempty, the equations of the type for k', k'' exist. So, k'' yields $\lambda > 0$, which ultimately leads to $\mu_{k'} > 0$ and $\alpha_{k'}^* = 0$. Similarly, $\mu_{k'''} > 0$ leading to $\alpha_{k'''}^* = 0$. This means that the only categories over which to search are in $B = (A \cap B) \cup (A^c \cap B)$. Over this smaller simplex, the α_i's over B sum to one, and only the categories in $A \cap B$ are active variables. To wit, the objective function over B reduces to,

$$\psi(\nu\alpha_{A\cap B} + n_{A\cap B}) - \psi(\nu + n_B).$$

Because ψ is monotonically increasing, this function achieves its minimum when $\alpha_{A\cap B}^* = 0$, such that $\alpha_{A^c\cap B}^* = 1$. In the general multivariate setting, any solution on the face of $\alpha_{A^c\cap B} = 1$ can attain the minimum value, so consists of all the minimisers of the problem. Note that this result continues to hold even if $A \cap B = \emptyset$, as the objective function over B becomes constant in this case,

$$-\psi(\nu + n_B).$$

□

5.1.4 Log odds ratios

Let us extend the analysis of the log-probability ratio to the log-odds case. Again, let L be a finite number of categories, $A, B \subseteq L$, $\nu > 0$ be the predefined hyperparameter of the IDM, $\boldsymbol{\theta} \in \triangle^L$ be a multinomial probability vector over L, and $\boldsymbol{n} \in \mathbb{N}^L$ be a vector of non-zero counts. We are interested in computing the following quantity,

$$\begin{aligned}\underline{E}_{\mathrm{IDM}}\left(\log\frac{\theta_A(1-\theta_B)}{(1-\theta_A)\theta_B}\,\middle|\,\nu,\boldsymbol{n}\right)\\ =\psi(\nu\alpha_A+n_A)+\psi(\nu\alpha_{B^c}+n_{B^c})\\ -\psi(\nu\alpha_B+n_B)-\psi(\nu\alpha_{A^c}+n_{A^c}).\end{aligned}$$

We can partition the possible KKT stationarity conditions for this problem as,

$$\begin{aligned}\forall k\in A\cup B^c,\ k\in A^c\cup B:\ &\nu\psi'_A[k\in A]+\nu\psi'_{B^c}[k\in B^c]-\nu\psi'_{A^c}[k\in A^c]-\nu\psi'_B[k\in B]-\mu_k+\lambda=0,\\ \forall k'\in A\cup B^c,\ k'\notin A^c\cup B:\ &\nu\psi'_A[k'\in A]+\nu\psi'_{B^c}[k'\in B^c]-\mu_{k'}+\lambda=0,\\ \forall k''\notin A\cup B^c,\ k''\in A^c\cup B:\ &-\nu\psi'_{A^c}[k''\in A^c]-\nu\psi'_B[k''\in B]-\mu_{k''}+\lambda=0,\\ \forall k'''\notin A\cup B^c,\ k'''\in A^c\cup B:\ &-\mu_{k'''}+\lambda=0.\end{aligned}$$

Once again, when the corresponding equations exist, we can deduce the following algebraic consequences of each of the equations.

$$\begin{aligned}\forall k'\in A\cup B^c,\ k'\notin A^c\cup B:\ \mu_{k'}>\lambda.\\ \forall k''\notin A\cup B^c,\ k''\in A^c\cup B:\ \lambda>0.\\ \forall k'''\notin A\cup B^c,\ k'''\notin A^c\cup B:\ \mu_{k'''}=\lambda,\end{aligned}$$

Let us again give the problem more structure and information to analyse the two classes k and k''.

Example 5.1.2: Suppose $A=\{1,2\}$ and $B=\{2,3\}$ with $L=\{1,2,3,4\}$ and fix a hyperparameter ν and dataset $\boldsymbol{n}$. Now,

$$\begin{aligned}k\in(A\cup B^c)\cap(A^c\cup B)=\{2,4\},\\ k'\in(A\cup B^c)\cap(A^c\cup B)^c=\{1\},\\ k''\in(A\cup B^c)^c\cap(A^c\cup B)=\{3\},\\ k'''\in(A\cup B^c)^c\cap(A^c\cup B)^c=\emptyset,\end{aligned}$$

There exist no $\alpha_{k'''}$'s in this problem. Because k' and k'' exist, we know that

$$\lambda>0,$$

and,

$$\mu_{k'} > \lambda > 0.$$

In this case, we obtain

$$\lambda^* > 0,$$

and

$$\alpha_1^* = 0.$$

To proceed with the equations of k, we consider the difference of trigamma functions therein under the condition that $\alpha_2 + \alpha_3 + \alpha_4 = 1$, a necessary condition for minimisation:

$$d_2 : (\alpha_2, \alpha_3, \alpha_4) \in \triangle^3 \longmapsto \psi'(\nu\alpha_2 + n_1 + n_2) - \psi'(\nu(\alpha_2 + \alpha_3) + n_2 + n_3),$$

and,

$$d_4 : (\alpha_2, \alpha_3, \alpha_4) \in \triangle^3 \longmapsto \psi'(\nu\alpha_4 + n_1 + n_4) - \psi'(\nu(\alpha_3 + \alpha_4) + n_3 + n_4).$$

Knowing $\lambda^* > 0$, there are cases of the signs of d_2 and d_4 that allow us to identify more zeroes from the complementary conditions $\mu_i\alpha_i^* = 0$. For example, if $\boldsymbol{n}$ has been observed such that $d_2(\cdot|\boldsymbol{n}, \nu)$ and $d_4(\cdot|\boldsymbol{n}, \nu)$ are nonnegative functions over $\boldsymbol{\alpha}$, then we can respectively show that,

$$d_2 \geq 0 \;\Rightarrow\; 0 < \lambda \leq \mu_2 \;\Rightarrow\; \alpha_2^* = 0,$$

and

$$d_4 \geq 0 \;\Rightarrow\; 0 < \lambda \leq \mu_4 \;\Rightarrow\; \alpha_4^* = 0.$$

We can solve for these conditions as inequalities by using the decreasing monotonicity of ψ' over positive arguments:

$$d_2 \geq 0 \iff \alpha_3 \geq \frac{1}{\nu}(n_1 - n_3),$$

and,

$$d_4 \geq 0 \iff \alpha_3 \geq \frac{1}{\nu}(n_1 - n_3).$$

If optimisation problem is such that $\alpha_3^* \geq (n_1 - n_3)/\nu$, then, knowing $\lambda^* > 0$, we can conclude,

$$\mu_2^* = \lambda^* + d_2(\boldsymbol{\alpha}^*) > 0 \;\Rightarrow \alpha_2^* = 0,$$

and,

$$\mu_4^* = \lambda^* + d_4(\boldsymbol{\alpha}^*) > 0 \;\Rightarrow \alpha_4^* = 0.$$

Because we have already identified $\alpha_1^* = 0$ earlier, when this happens, we can directly conclude that $\alpha_3^* = 1$.

We briefly remark that this line of reasoning depends on the fact that $(n_1 - n_3)/\mu \leq 1$: there is no information on the positivity of d_2 and d_4 if this does not hold. From this we can deduce the following effect of $\boldsymbol{n}$ and ν on this particular example: larger values of ν and smaller (but nonnegative) *differences* of $n_1 - n_3$ causes the optima to localise to a small region of the simplex (namely, the singleton $\{(0, 0, 1, 0)\} = \{\boldsymbol{\alpha} \in \overline{\triangle^4} : \ \alpha_3 = 1\}$).

■

5.1.5 Independence test statistic

Relative to the log-probability and log-odds ratios studied above, the independence test statistic $\theta_{A_1} \dots \theta_{A_r} / \theta_{A_1 \cap \dots \cap A_r}$ has a significant structure regarding the posterior distributions of the IDM. Let us demonstrate these with the KKT conditions of inference for the independence statistic of a cell in an $I \times J$ contingency table.

Consider $L = \{uv : \ u = 1, \dots, I, \ v = 1, \dots, J\}$ with $I, J < \infty$. For a particular cell, ij, we are interested in computing the posterior lower expectation of,

$$\underline{E}_{\text{IDM}} \left(\log \frac{\theta_{i\cdot} \theta_{\cdot j}}{\theta_{ij}} \,\middle|\, \boldsymbol{n}, \nu \right) = \psi(\nu\alpha_{i\cdot} + n_{i\cdot}) + \psi(\nu\alpha_{\cdot j} + n_{\cdot j}) - \psi(\nu\alpha_{ij} + n_{ij}) - \psi(\nu + n),$$

where, for example, $\alpha_{i\cdot} = \sum_{v=1,\dots J} \alpha_{iv}$.

Notice that, denominator categories, $\{ij\}$ is nested in the numerator categories $\{i\cdot\} \cup \{\cdot j\}$. This means that there is only one category that is in both the numerator and denominator sets, and there are no categories that are only in the denominator but not the numerator. So, the possible KKT equations are,

$$\begin{aligned}
k = ij : \ & \nu\psi'_{i\cdot} + \nu\psi'_{\cdot j} - \nu\psi'_{ij} - \mu_{ij} + \lambda = 0, \\
k' \in \{i\cdot\} \cup \{\cdot j\}, \ k' \neq ij : \ & \nu\psi'_{i\cdot}[k' \in \{i\cdot\}] + \nu\psi'_{\cdot j}[k' \in \{\cdot j\}] - \mu_{k'} + \lambda = 0, \\
k'' \notin \{i\cdot\} \cup \{\cdot j\}, \ k'' = ij : \ & \text{(does not exist)}, \\
k''' \notin \{i\cdot\} \cup \{\cdot j\}, \ k'' \neq ij : \ & \mu_{k'''} - \lambda = 0.
\end{aligned}$$

Now, given a contingency table, the three classes of equations always exist since they partition the table. Then, we have,

$$\mu_{k'} > \lambda,$$

and

$$\mu_{k'''} = \lambda.$$

Note that this implies,

$$\lambda \geq 0,$$

so that, $\mu_{k'} > 0$ and,

$$\alpha^*_{k'} = 0.$$

That is, all $\alpha^*_u = 0$, where its cell u is on either the i-th row or the j-th column, but not equal to ij itself. Finally, note that there is only a single equation in the k class: its difference of trigamma functions is,

$$d_{ij}(\boldsymbol{\alpha}) = \nu\psi'_{i\cdot}(\boldsymbol{\alpha}) + \nu\psi'_{\cdot j}(\boldsymbol{\alpha}) - \nu\psi'_{ij}(\boldsymbol{\alpha}).$$

As we have observed in past examples, for some values of $\boldsymbol{n}$ and ν, it is possible for the differences of the first equation class can have a constant sign over $\triangle^{IJ} \ni \boldsymbol{\alpha}$.

Example 5.1.3: Let us study the sign of d_{ij} with a 2×2 table. Suppose $I, J = 2$ and $ij = 11$. Then

$$\psi'(\nu(\alpha_{11} + \alpha_{12}) + n_{11} + n_{12}) + \psi'(\nu(\alpha_{11} + \alpha_{21}) + n_{11} + n_{21}) - \psi'(\nu\alpha_{11} + n_{11}),$$

over $0 \leq \alpha_{11} + \alpha_{12} + \alpha_{22} \leq 1$. But, we know that $\alpha^*_{12}, \alpha^*_{21} = 0$, so the domain of interest of the difference function simply becomes $\alpha_{11} \in [0, 1]$ (with $\alpha_{22} = 1 - \alpha_{11}$, and so we are interested in,

$$d_{11}(\alpha_{11}) = \psi'(\nu\alpha_{11} + n_{11} + n_{12}) + \psi'(\nu\alpha_{11} + n_{11} + n_{21}) - \psi'(\nu\alpha_{11} + n_{11}).$$

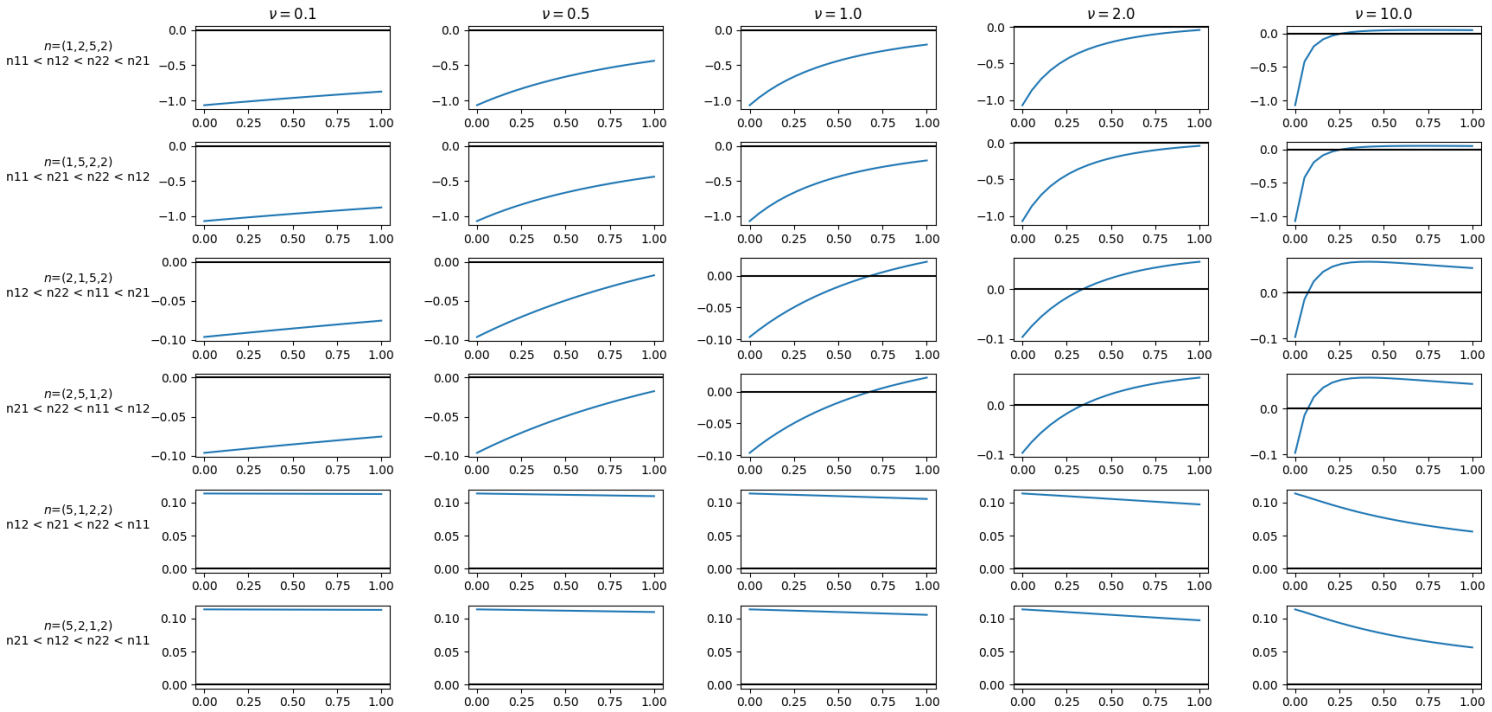

Figure 5.1: The difference function, $\alpha_{11} \longmapsto \psi'(\nu\alpha_{11} + n_{11} + n_{12}) + \psi'(\nu\alpha_{11} + n_{11} + n_{21}) - \psi'(\nu\alpha_{11} + n_{11})$, plotted over $\alpha_{11} \in [0, 1]$ for various datasets $\boldsymbol{n} = (n_{11}, n_{12}, n_{21}, n_{22})$ that total to 10 and permute n_1, n_2, n_3 over $\{1, 2, 5\}$, and various ν values in $\{0.1, 0.5, 1, 2, 10\}$.

Figure 5.1 plots this univariate function for various values of $\boldsymbol{n}$ and ν. The figure shows that there are certain values of $\boldsymbol{n}$ and ν for which $d_{11}(\alpha_{11}) > 0$, $d_{11}(\alpha_{11}) \leq 0$ and for which $d_{11}(\alpha_{11})$ changes sign over $\alpha_{11} \in [0, 1]$. In the first case, $d_{11} > 0$ over $[0, 1] \ni \alpha_{11}$ implies that $d_{11}(\alpha_{11}^*) \geq 0$, leading to,

$$0 < d_{11}(\alpha_{11}^*) = \mu_{11}^* - \lambda^* = 0,$$

which, along with $\lambda \geq 0$, implies $\mu_{11}^* > 0$ and so $\alpha_{11}^* = 0$. In the second and third cases, we cannot use the complementary conditions to conclude any additional zeroes in $\boldsymbol{\alpha}^*$. However, the fact that d_{11} is univariate can be exploited more directly as follows. Notice

that d_{11} is in fact the derivative of the objective function restricted to $\alpha_{12}, \alpha_{21} = 0$,

$$d_{11}(a) = \frac{d}{da}(\psi(\nu a + n_{11} + n_{12}) + \psi(\nu a + n_{11} + n_{21}) - \psi(\nu a + n_{11})).$$

This means that, if d_{11} is negative over $[0, 1]$ as in the second case, then the restricted objective function is a non-increasing function over $[0, 1]$, which must achieve its minimum at $\alpha_{11}^* = 1$. In the third case where $d_{11}(\alpha_{11})$ is continuous, monotonically non-decreasing on $[0, 1]$ and has a root at, say $a_0 \in [0, 1]$, then it means that a_0 is a stationary point of the restricted objective function that is decreasing in $[0, a_0)$ and increasing in $(a_0, a]$. Because the restricted objective is continuous $[0, 1]$, then a_0 must be its minimum, such that $\alpha_{11}^* = a_0$. This characterises all the possible cases.

■

5.2 Some properties of the objective function in mean-parameter space

It is expected that one would numerically optimise the objective function over the subspace identified by the KKT conditions in the first step, and this subspace is another closed simplex of lower dimension than that with which one started. Towards an analysis of this optimisation, it remains unclear that the objective function itself has any monotonic or convexity properties.

We show that there exists a reparametrisation of the problem such that the convexity of the domain in the natural parameter space is preserved, the posterior log-odds expectation to be optimised is bounded in the interior of the space under the reparametrisation and that there are some regions of the new space over which the objective function is locally monotone. The latter is useful as we will show that some datasets and log-odds structures restrict the optimisation of the posterior expectation to this region, making the function monotone over this set.

5.2.1 A reparametrisation of the natural parameter space

For illustration, we graphically explore the salient ideas with a low dimensional problems instead of deriving the general finite dimensional case. Furthermore, the parametrisation

is simply a tool to transform the problem into one that is more amenable to a gradient descent, and it is neither unique nor has any other significance. We will therefore focus on an illustration with an example for the rest of the section.

Example 5.2.1: Given a dataset of counts $\boldsymbol{n}$ over categories $L = \{1, 2\}$ totaling to n and a hyperparameter choice $\nu > 0$, the posterior IDM lower expectation of a log-odds statistic T, is the minimum of,

$$I : \boldsymbol{\alpha} \in \triangle^2 \longmapsto \psi(\nu\alpha_1 + n_1) - \psi(\nu\alpha_2 + n_2),$$

where we note that $\alpha_2 = 1 - \alpha_1$ and $n_2 = n - n_1$. Recall that the mean parameters are given by,

$$\mu_i = \psi(\nu\alpha_i + n_i),$$

for $i = 1, 2$, with the constraint that,

$$\psi^{-1}(\mu_1) + \psi^{-1}(\mu_2) = \nu + n. \tag{5.1}$$

We will study (5.1) in the subsequent set of examples via a convenient reparametrisation. The parametrisation allows us to derive some of the curve's properties without resorting to direct manipulation (5.1), as well as reveals interesting geometrical properties of the optimisation problem.

Definition 5.2.1: (Reparametrisation of $\boldsymbol{\mu}$ curve) Define the parameter $t \in \mathbb{R}$, and,

$$\boldsymbol{\mu}(t) = -a(t)\boldsymbol{v}_0 + t\boldsymbol{v}_1,$$

such that $\boldsymbol{v}_0$, $\boldsymbol{v}_1$ and $a(t)$ are chosen to satisfy,

$$\psi^{-1}(a(t)v_{01} + tv_{11}) + \psi^{-1}(a(t)v_{02} + tv_{12}) = \nu + n.$$

□

We can visualise this parametrisation in Figure 5.2.

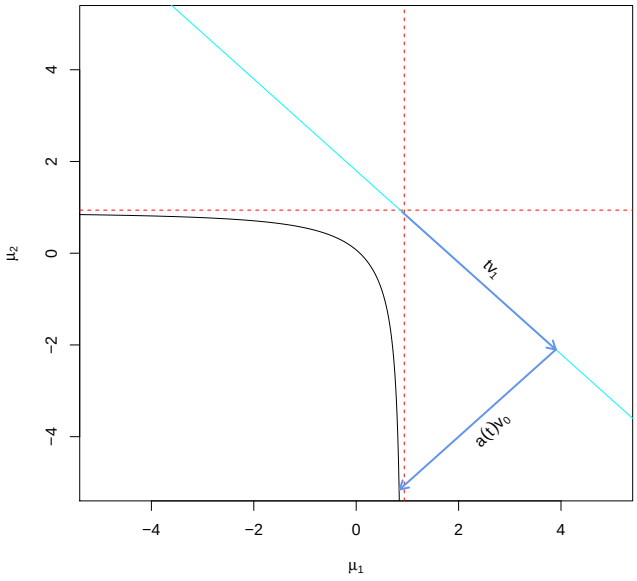

Figure 5.2: A visualisation of the parametrisation $\boldsymbol{\mu}(t) = a(t)\boldsymbol{v}_0 + t\boldsymbol{v}_1$ for two mean parameters. The black curve is the set of points satisfying $\psi^{-1}(\mu_1) + \psi^{-1}(\mu_2) = \nu$, the normalisation constraint of the natural parameters of the Dirichlet distribution. The red dashed lines are the asymptotes of the black curve. The cyan line is the hyperplane onto which $\boldsymbol{\mu}(t)$ is projected bijectively.

The map $\boldsymbol{\mu}(t)$ takes t, maps it to a point on the linear subspace $t\boldsymbol{v}_1$, then translates it by $-a(t)\boldsymbol{v}_0$ to a point on the constrained curve. The particular choice of $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ is not important: it suffices that $\boldsymbol{v}_0, \boldsymbol{v}_1$ and $a(t)$ place the subspace to one side of the curve $\boldsymbol{\mu}$. We have, for convenience, chosen $\boldsymbol{v}_0$ to be the normal vector and $\boldsymbol{v}_1$ to be orthogonal to $\boldsymbol{v}_0$ and such that the subspace spanned by $\boldsymbol{v}_1$ intersects the origin. The mapping $t \longmapsto \boldsymbol{\mu}(t)$ is therefore bijective when $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ are chosen in such a way.

■

Example 5.2.2: (Convexity of the epigraph of $\boldsymbol{\mu}(t)$ and $-a(t)$) From Figure 5.2, given that we fix the choice of the hyperplane, then $-a(t)$ traces out the curve $(\mu_1(t), \mu_2(t))$. Furthermore, it is clear that the epigraph of $\boldsymbol{\mu}$ is a convex region and, equivalently, the

outline of this epigraph, $-a(t)$, is a convex function.

■

Example 5.2.3: (Reparametrised images of faces) It is important to understand how the faces of the simplex of the natural parameters map to the mean parameter space. Not only does this give a complete picture of the transformed domain, but, as in examples like the log-odds, the boundaries may contain optima.

Continuing with Example 5.2.1, the (singleton) faces $\alpha_i = 0$ map to $\mu_i = \psi(n_i)$ in the mean parameter space, for $i = 1, 2$. In particular, if $n_i = 0$, then $\alpha_i = 0$ maps to an infinity in $\boldsymbol{\mu}$ space. To put this in terms of t, the (singleton) face $\alpha_1 = 0$ is mapped to the set,

$$\{t : \ a(t)\boldsymbol{v}_0 + t\boldsymbol{v}_1 = (\psi(n_1), \psi(\nu + n_2))\}.$$

It follows from the bijectivity we have established between t and $\boldsymbol{\mu}(t)$ that this is a singleton set that maps to a single vector $(\psi(n_1), \psi(\nu + n_2))$. We can similarly establish the image of the singleton face of $\alpha_2 = 0$ in terms of t.

We will see more nontrivial images of faces of $\triangle^3$ in subsequent examples.

■

5.2.2 Geometry of the objective function

Having established the necessary concepts in two-dimensions, we can explore the non-trivial geometry that arises in the three-dimensional case that is also easily visualised.

Example 5.2.4: Suppose that we have a dataset of counts $\boldsymbol{n}$ over categories $L = \{1, 2, 3\}$ totaling to n and a hyperparameter choice $\nu > 0$. Let us consider the log odds statistic,

$$\log \frac{\theta_1 + \theta_2}{\theta_2 + \theta_3},$$

whose posterior IDM lower expectation is the minimum of,

$$I : \boldsymbol{\alpha} \in \triangle^3 \longmapsto \psi(\nu(\alpha_1 + \alpha_2) + n_1 + n_2) - \psi(\nu(\alpha_2 + \alpha_3) + n_2 + n_3).$$

Again, the mean parameters are given by,

$$\mu_i = \psi(\nu\alpha_i + n_i),$$

for $i = 1, 2, 3$, with the constraint that,

$$\psi^{-1}(\mu_1) + \psi^{-1}(\mu_2) + \psi^{-1}(\mu_3) = \nu + n.$$

Definition 5.2.2: Define $\boldsymbol{t} = (t_1, t_2)$, and

$$\boldsymbol{\mu}(\boldsymbol{t}) = -a(t_1, t_2)\boldsymbol{v}_0 + t_1\boldsymbol{v}_1 + t_2\boldsymbol{v}_2,$$

such that $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $a(t_1, t_2)$ are chosen to satisfy,

$$\psi^{-1}(\mu_1(\boldsymbol{t})) + \psi^{-1}(\mu_2(\boldsymbol{t})) + \psi^{-1}(\mu_3(\boldsymbol{t})) = \nu + n.$$

□

The convexity of the epigraph of $\boldsymbol{\mu}(\boldsymbol{t})$ is given by the fact that we choose $\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2$ to be linearly independent and, for convenience, orthogonal. Also, Lemma D.1.4 asserts that $t_1, t_2 \longmapsto -a(t_1, t_2)$ is a convex function. Then, the mapping from $(t_1, t_2, -a(t_1, t_2))$ to $(\mu_1(\boldsymbol{t}), \mu_2(\boldsymbol{t}), \mu_3(\boldsymbol{t}))$ is simply an invertible rotation of the constrained surface $\boldsymbol{\mu}$, and so $\boldsymbol{\mu}(\boldsymbol{t})$ preserves the convexity of the graph $(t_1, t_2, -a(t_1, t_2))$.

The face of the natural parameter simplex $\triangle^3$ where, say $\alpha_i = 0$ for all i's in some index set I is mapped to sets of $\boldsymbol{t}$'s as,

$$\{(t_1, t_2) : \ a(t_1, t_2)\boldsymbol{v}_{0i} + t_1\boldsymbol{v}_{1i} + t_2\boldsymbol{v}_{2i} = \psi(n_i) \text{ for all } i \in I\}.$$

Notice that $n_i = 0$ (for example, when the data is sparse, or when performing prior inference) causes some of the components of $\boldsymbol{\mu}(\boldsymbol{t})$ to become unbounded in certain directions. For this example, $\mu_1 = \psi(n_1)$ implies that $\alpha_1 = 0$ and $\alpha_2 + \alpha_3 = 1$, so the face for $\alpha_1 = 0$ is mapped to,

$$\{(t_1, t_2) : \ \mu_1 = a(t_1, t_2)\boldsymbol{v}_{01} + t_1\boldsymbol{v}_{11} + t_2\boldsymbol{v}_{21} = \psi(n_1)\},$$

and the other two faces map similarly. The contours,

$$\{(t_1, t_2) : \ \mu_1(t_1, t_2) = c_1\}, \{(t_1, t_2) : \ \mu_2(t_1, t_2) = c_2\}, \{(t_1, t_2) : \ \mu_3(t_1, t_2) = c_3\}$$

for fixing μ_1, μ_2 and μ_3 to various respective real values, c_1, c_2, c_3, are plotted in Figure 5.3.

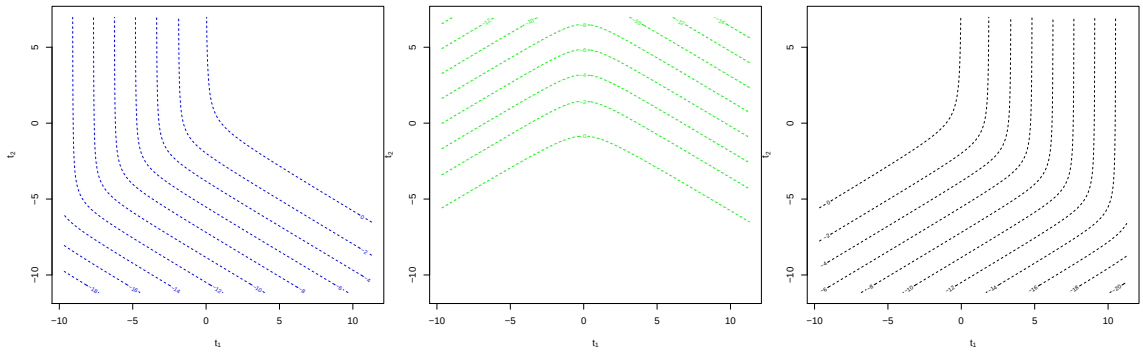

Figure 5.3: Level curves of $\mu_i(t_1, t_2) = a(t_1, t_2)\boldsymbol{v}_{0i} + t_1\boldsymbol{v}_{1i} + t_2\boldsymbol{v}_{2i}$ over t_1, t_2 space for when μ_1, μ_2 and μ_3 being held constant (respectively, left, middle and right).

Importantly, the resulting optimisation domain in the t_1-t_2-space corresponds to the area trapped between these curves. For example, the three dimensional simplex, consisting of the three edges $\alpha_1 = 0$, $\alpha_2 = 0$, $\alpha_3 = 0$, will, pick out the corresponding contours,

$$\mu_1(t_1, t_2) = \psi(n_1),\ \mu_2(t_1, t_2) = \psi(n_2),\ \mu_3(t_1, t_2) = \psi(n_3),$$

as boundaries of the optimisation domain in this space. Finally, these contours move to the infinities of $\mathbb{R}^2$ as n_i tends to zero.

■

Example 5.2.5: (Geometry of the objective function) The objective function can be written in terms of t_1 and t_2:

$$t_1, t_2 \longmapsto \psi(\nu(\alpha_1(\boldsymbol{\mu}(t_1, t_2))+\alpha_2(\boldsymbol{\mu}(t_1, t_2)))+n_1+n_2)-\psi(\nu(\alpha_2(\boldsymbol{\mu}(t_1, t_2))+\alpha_3(\boldsymbol{\mu}(t_1, t_2)))+n_2+n_3),$$

where $\alpha_i(\boldsymbol{\mu})$ transforms the mean parameter back to the i-th natural parameter. As noted before, the domain of optimisation is unbounded when the n_i's are zero. When all of them are zero, it corresponds to the optimisation problem of the prior lower expectation. The objective function's contours in this case are visualised in Figure 5.4.

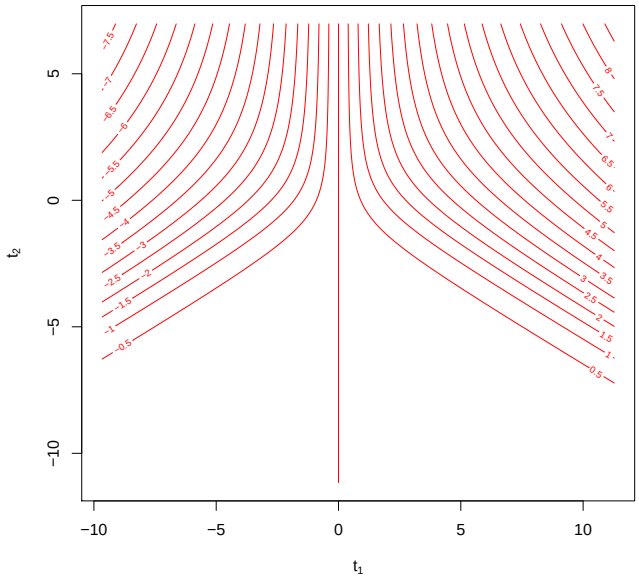

Figure 5.4: Contours of the objective function as a function of t_1, t_2 when $n = 0$.

We can observe that there is a saddle point line at $t_1 = 0$, and the function is increasing as one heads left away from the line in certain directions, and decreasing on the right of the line similarly. The function is also unbounded in these regions. This suggests that the optima of the prior lower expectation problem are also unbounded in $\mathbb{R}^2$. By Lemma D.1.3, in the $\boldsymbol{\mu}$ parametrisation, the objective function has no turning points in the interior of $\mathbb{R}^3 \ni \boldsymbol{\mu}$.

As noted in Example 5.2.4, when data is observed, the optimisation domain becomes restricted by the constant contours of $\boldsymbol{\mu}(t_1, t_2)$. This is shown in Figure 5.5, where the objective function contours of Figure 5.4 are overlayed and restricted by the contours in Figure 5.3 under various cases of the data counts n_1, n_2 and/or n_3 being zero. In Figure 5.5, we see that if certain n_i's have zero counts, then that boundary moves to infinity in its corresponding direction.

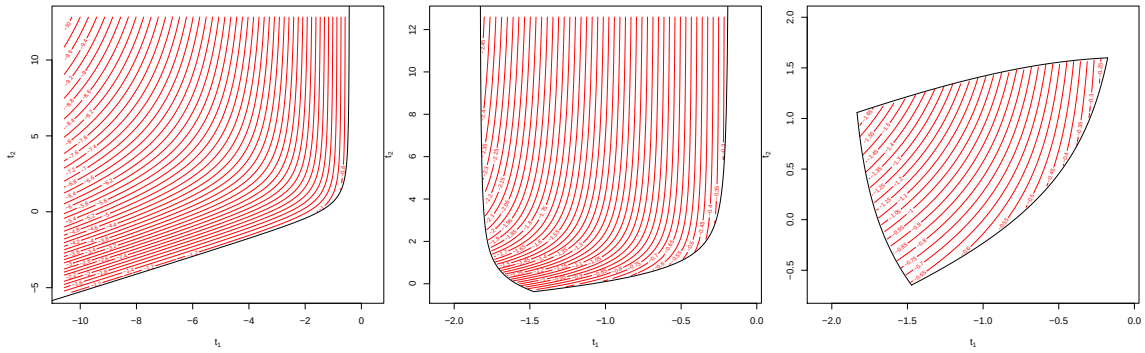

Figure 5.5: Various restrictions of the domain of optimisation with the objective function's contours being plotted only inside this domain. The left panel is when $n_1 = n_2 = 0$ such that only the points of the boundary satisfying $\mu_3(\boldsymbol{t}) = \psi(n_3)$ with $n_3 > 0$ are finite, and the other two boundaries tend to the infinities of the t_1, t_2 space. The middle panel is when $n_2 = 0$ only and the right panel is when $n_1, n_2, n_3 > 0$.

We observe graphically that indeed the optimisation domains are convex. Furthermore, as it happens, certain datasets place the domain strictly on one side of the saddle point line, making the objective function monotone within the domain of optimisation. For example, the contours of the right panel in Figure 5.5 restrict the posterior objective function to the left of saddlepoint region $t_1 = 0$ in of the unrestricted objective function in Figure 5.4. Finally, regions where the objective function is unbounded only happens at the corresponding boundary of the domain at the infinities of the t_1-t_2 space. This demonstrates that the t_1, t_2 parametrisation yields properties that make the problem readily amenable to gradient descent optimisation.

■

5.3 Concluding remarks

Our contributions in this chapter are as follows. We use Lemmas 5.1.1, 5.1.2 and 5.1.3 to aid in the analysis of the Karush-Kuhn-Tucker condition of the optimisation problems of the IDM lower expectations of three common log-odds statistics: the log-probability (Example

5.1.1 and Theorem 5.1.1), the log-odds (Example 5.1.2) and the independence statistic (Example 5.1.3). These demonstrate how the KKT conditions can narrow down the location of the optimisers to certain faces of the natural parameters simplex. To geometrically study the objective function, we construct a reparametrisation of the optimisation problem explore the geometry of the objective function in the mean parameter space: this is illustrated conceptually in two dimensions in Definition 5.2.1, Examples 5.2.1, 5.2.2 and 5.2.3, and salient properties of higher dimensions are illustrated in three dimensions Examples 5.2.4 and 5.2.5.

In this chapter the optimisation problem involved in computing the posterior lower expectation of log-odds statistics under the IDM was explored. For three commonly occurring log-odds statistics, their KKT conditions were solved and used to identify the faces of the natural parameter space containing the optimisers. Considerations were given to gradient based methods typically used to search for the optimisers on these faces. In particular, we have noted that these methods are problematic when searching for optimisers of unbounded objective functions over a closed set such as those occurring with the IDM log-odds problems. In this vein, we moved beyond the KKT condition to study the shape of the objective function in the mean parameter space where the region over which the optimiser is unbounded is at the infinities of the Euclidean space of mean parameters. A reparametrisation via a projection was developed to illustrate that domain of optimisation is convex in both prior and posterior cases and, importantly, to provide a proof of concept that the objective function can be monotone for certain log-odds structure and parametrised by certain datasets.

The exploration done in this work suggests that, while the original optimisation problem of the IDM lower expectation problem is not a convex optimisation problem, it possesses properties that nevertheless make it amenable to the usual tools used to analyse optimisation problems. From a statistical perspective, our study is pertinent regarding a global sensitivity analysis over the *closed* simplex of the natural parameters of a Dirichlet family of priors with fixed concentration parameter. Our studies suggest that, at least for log-odds statistics, such global analysis is amenable to optimisation techniques under the IDM due to the properties of the objective function and its domain. In fact, with log-odds of certain structures, the observed dataset may parametrise the optimisation problem as to provide a sufficient amount of information and discrimination for the KKT conditions to locate the set of optima by itself (for example when observations are predominantly or even completely observed only in the sets involved in the numerator or only in those of the denominator). Furthermore, as we have seen in Figures 5.4 and 5.5, the objective func-

tion, in both the mean parameter and its reparametrisation which we have constructed, may exhibit directions in which it is monotone, such that its optima always occur at the boundary of the domain of optimisation. This *geometrical* fact, which is not apparent from viewing the objective function in the natural parameters as a difference of digamma functions, sheds light on the nature of the IDM log-odds problem and, equivalently, the global sensitivity analysis of Dirichlet prior families and its feasibility.

As gradient-based methods are a typical choice for such problems, we briefly remark on the ramifications of our exploration upon its use. One may encounter two issues in applying them to an IDM log-odds minimisation. First, because the objective function is potentially an arbitrary sum and difference of digamma functions, it is unclear that it has nice properties such as monotonicity and convexity which the optimiser algorithm can exploit. Second, as we have illustrated above, certain structures of sets in the log-odds and datasets that parametrise the posterior problem yield vertex solutions whose objective values are unbounded, ill-defined and cannot be smoothly extended to the boundary of the optimisation domain. In this analysis, we have shown that there exists a reparametrisation of the problem such that the convexity of the domain is preserved, the posterior log-odds expectation to be optimised is bounded in the interior of the space under the reparametrisation and that there are some regions of the new space over which the objective function is locally monotone.

Chapter 6

Imprecise posterior inference under zero lower marginal probability in finite dimensions

In this chapter, we focus on an inferential issue that, in practice, does not occur when a single distribution is used. Recall that we are partially motivated by the case where a set of priors has been elicited from a set of experts, who may not hold consistent beliefs. The most extreme example of this would be when they differ in what they assess to have strictly positive probability. Hence, some distributions in an elicited prior set assign zero marginal probability to the possibly observed data. As a result, the lower probability (see Definition 2.1.9) of this particular dataset is zero, and the generalised Bayes' rule of Walley [81] (Theorem 2.2.4) does not necessarily hold. In these cases, *extensions* of the lower expectation may be used instead.

We explore whether or not coherence and imprecision might yield well-defined and sensible inference when the lower probability of the conditioning set is zero. This is motivated by the fact in, for example, Walley [81], Couso and Moral [29] and Quaeghbeur (Ch 1. of Augustin et al. [6]), that the same unconditional lower expectation model may generate different conditional imprecise models when the lower probability of the conditioning set is zero. Our finite discrete setting allows tractable and concrete exploration of the properties directly relevant to this issue.

We focus on two established imprecise methods that can be used in this situation. The

vacuous extension always yields vacuous inference. (Recall, from Definition 2.2.2, that imprecise expectations are vacuous for a random variable if the lower and upper expectations are respectively the infimum and supremum of the random variable.) On the other hand, under certain conditions, the *regular extension* may yield non-vacuous inference even when the marginal lower probability is zero. In this chapter, we discuss in detail a very simple example to gain insight into the characteristics and problems of statistical inference arising from these two extensions.

6.1 A running example

Let $\{1, \ldots, m\}$ index the known categories in consideration with $m < \infty$. We consider the case when the sample size n is apriori fixed and each observation is i.i.d. following some multinomial distribution. Let $\boldsymbol{n} \in \mathbb{N}^m$ be the counts of the categories summing to n. Let $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{|\Theta|}\} \subset \overline{\triangle^m}$ be a finite set of candidate multinomial probability vectors. A prior $\boldsymbol{p}$ over Θ is therefore a point on the finite dimensional simplex $\overline{\triangle^\Theta}$. When the denominator is non-zero, the posterior expectation of a random variable $f(\boldsymbol{\theta})$ following from Bayes' rule is given by,

$$E_{\boldsymbol{p}}(f) = \frac{\sum_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\boldsymbol{n}) \boldsymbol{p}(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\boldsymbol{n}) \boldsymbol{p}(\boldsymbol{\theta})}. \tag{6.1}$$

Given a set M of priors, $\boldsymbol{p}$'s, a posterior lower expectation can be formed by the optimisation of the posterior expectation over the set M,

$$\underline{E}(f|\boldsymbol{n}) = \inf \{E_{\boldsymbol{p}}(f(\boldsymbol{\theta})|\boldsymbol{n}) : \ \boldsymbol{p} \in M\}. \tag{6.2}$$

Recall from Definition 2.1.9, the lower probability of an event A under $\underline{E}$ is $\underline{P}(A) = \underline{E}(I_A)$. As a shorthand notation, we will write the lower probability of observing $\boldsymbol{n}$ as $\underline{P}(\boldsymbol{n})$.

When $\underline{P}(\boldsymbol{n}) > 0$, (6.2) is a coherent lower expectation as a consequence of the generalised Bayes' rule (Theorem 2.2.4). But, in this chapter, we are interested in the case when M is such that this is not the case. To illustrate the context of the need for imprecision under this finite setting, we furnish the following example.

Example 6.1.1: A linguist wishes to model the probability $f(\theta) = \theta$, $\theta \in [0, 1]$, of a certain phrase occurring in a corpus of documents all written in a single language. The phrase is grammatically incorrect in the documents' language: it is typical to assign a

structural zero to the occurrence of phrases which are syntactically not permissible (Mohri and Roark [61]). That is, $\theta_1 = 0$ is a candidate model. However, it is also possible that such a document is written in a way that the phrase nevertheless occurs: for example, the phrase may be used as a colloquialism or written by somebody unfamiliar with the language. To account for these cases, the linguist selects two other distributions, $\theta_2, \theta_3 > 0$. In all, the candidate distributions for the document are $\Theta = \{\theta_1, \theta_2, \theta_3\}$.

The linguist wishes to construct a prior distribution over Θ. Because the linguist's prior state of knowledge does not favour any particular distribution, any prior distribution (assigning any probability to θ_1) is consistent with this state. In particular, this includes the prior that assigns a prior probability of one to $\theta_1 = 0$ which will cause the marginal probability of observing a document containing the phrase to be zero. That is, writing n_1 to be the number of documents containing this phrase out of n observed documents,

$$\begin{aligned}
&P(n_1 \text{ documents of } n \text{ contain the phrase}) \\
&\quad = P(\boldsymbol{n} = (n_1, n - n_1)) \\
&\quad = p(\theta_1)L(\theta_1|(n_1, n - n_1)) + p(\theta_2)L(\theta_2|(n_1, n - n_1)) + p(\theta_3)L(\theta_3|(n_1, n - n_1)) \\
&\quad = 1 \cdot 0 + 0 \cdot L(\theta_2|(n_1, n - n_1)) + 0 \cdot L(\theta_3|(n_1, n - n_1)) \\
&\quad = 0.
\end{aligned}$$

This prevents any posterior inference involving conditional probabilities such as (6.1) to be well-defined.

■

This example shows that M may happen to be elicited in a way such that the denominator of (6.1) is zero for at least one potentially observable dataset. For such prior distributions in the set, we highlight two practical concerns regarding this.

Firstly, the posterior expectation (6.1) is not defined when its denominator is zero. In turn, this is so if and only if for every θ either its likelihood relative to the data or its prior probability (or both) are zero. What should be done here methodologically? On the one hand, one should not alter the prior after observing the data. A satisfactorily elicited prior should not be modified without any justification. On the other hand, one might apriori assume a prior that assigns a positive probability to any potential datasets. However, this prior may contradict any prior information that indicates that certain datasets in fact have

zero probability of occurring.

Secondly, when a dataset is observed such that its lower prior probability is zero, the generalised Bayes' rule does not guarantee a coherent posterior model. Moreover, under the global sensitivity analysis interpretation, the computation of the posterior lower expectation found by minimising (6.1) over the prior space is not well-defined in the sense that some Bayes' rules will not be defined with priors that produce a zero marginal probability for the data. As such, one cannot readily depend on the generalised Bayes' rule to generate posterior values that are coherent with the elicited set of priors.

From these observations, we do not expect that the lower and upper expectations of the generalised Bayes' rule to be readily applicable here, as it is a function of all the precise Bayes' rules over an elicited prior set. However, there are other more general ways of constructing imprecise posterior inferences beyond lower and upper bounds of conditional expectations.

6.2 Vacuous and regular extensions

Let us introduce some additional notations for this chapter. Consider a set of multinomial probability vectors over m categories, $\Theta \subset \overline{\triangle^m}$ such that $|\Theta| < \infty$, and M, set of prior distributions over Θ. Let $\boldsymbol{n}$ be a fixed dataset. For a prior,

$$\boldsymbol{p} = (p(\boldsymbol{\theta}_1), \ldots, p(\boldsymbol{\theta}_{|\Theta|})) \in \overline{\triangle^{|\Theta|}},$$

and likelihood,

$$\boldsymbol{L}(\boldsymbol{\theta}|\boldsymbol{n}) = (L(\boldsymbol{\theta}_1|\boldsymbol{n}), \ldots, L(\boldsymbol{\theta}_{|\Theta|}|\boldsymbol{n})).$$

where,

$$L(\boldsymbol{\theta}_i|\boldsymbol{n}) = \prod_{j=1}^{|\Theta|} \theta_{ij}^{n_j}.$$

Example 6.2.1: With only limited information about the source of the documents, the linguist decides to construct a set of candidate prior distributions $\boldsymbol{p} = (p(\theta_1), p(\theta_2), p(\theta_3)) = (p_1, p_2, p_3)$, M, that reflects the available but incomplete information. Suppose that the linguist elicits $\Theta = \{0, 1/10, 1/4\}$ as the possible probabilities of the phrase occurring and

that there are 10 documents in the sampled corpus. Suppose also that from prior information of similar documents, the linguist assesses an upper bound on the prior expectation of odds of the phrase occurring, $\theta/(1-\theta)$, to be less than $1/4$,

$$0 \cdot p_1 + \frac{1}{9}p_2 + \frac{1}{3}p_3 \leq \frac{1}{4},$$

as well an upper bound on the prior expectation of the number of occurrences of the phrase in a document of size 10, $10 \cdot \theta$,

$$10 \cdot 0 \cdot p_1 + 10 \cdot \frac{1}{10}p_2 + 10 \cdot \frac{1}{4}p_3 \leq 2.$$

Now, M is to be constructed such that it reflects this known information. First, the linguist notes the lack of information relevant to assessing probability of a document following the rules of the language translates into including any $\boldsymbol{p}$ satisfying,

$$0 \leq p_1 \leq 1,$$

as a candidate. After some manipulation, the linguist's set of prior models is,

$$\left\{ (p_1, p_2, p_3) : \ (p_1, p_2, p_3) \in \overline{\triangle^3}, \ \frac{1}{9}p_2 + \frac{1}{3}p_3 \leq \frac{1}{4}, \ p_2 + \frac{5}{2}p_3 \leq 2 \right\}.$$

This set is a convex subset of the probability simplex in Figure 6.1, with vertices,

$$\{(1,0,0), (0,1,0), (0,3/8,5/8), (1/4,0/3/4)\}.$$

Importantly, note that $\boldsymbol{p} = (1,0,0)$, the prior assigning zero marginal probability to observing any document containing the phrase of interest, is an extreme point of this convex set.

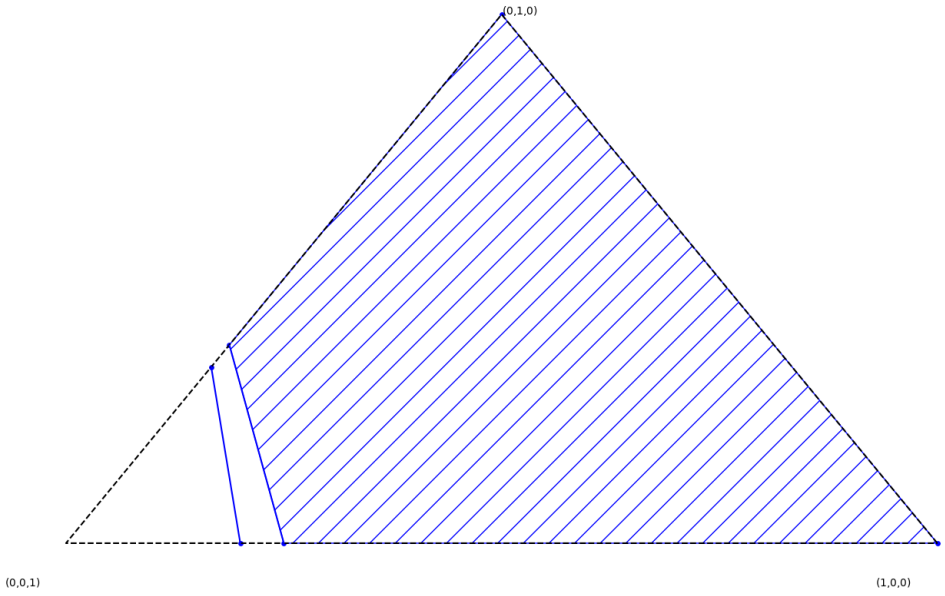

Figure 6.1: The linguist's set of priors based on two constraints (blue hatched). Notice that one of the constraints is redundant in defining the set. Notice also that $(1, 0, 0)$, the prior assigning zero marginal probabilities to certain datasets, is included in the set of priors.

■

The following example demonstrates the effects of the prior assigning a zero lower probability to a particular observed dataset under M when this dataset becomes observed.

Example 6.2.2: Consider a number of documents with the phrase is $n_1 \geq 1$ out of the sample of $n < \infty$, so the binomial count vector is $\boldsymbol{n} = (n_1, n - n_1)$. The likelihood vector becomes,

$$\boldsymbol{L} = (0, 0.1^{n_1} \cdot 0.9^{n-n_1}, 0.25^{n_1} \cdot 0.75^{n-n_1}).$$

Of all the priors in M, only $\boldsymbol{p} = (1, 0, 0)$ assigns zero marginal probability to $\boldsymbol{n} = (n_1, n - n_1)$. Indeed,

$$\boldsymbol{L} \cdot \boldsymbol{p} = (0, 0.1^{n_1} \cdot 0.9^{n-n_1}, 0.25^{n_1} \cdot 0.75^{n-n_1}) \cdot (1, 0, 0) = 0, \tag{6.3}$$

and by the non-negativity of all the values involved, $\boldsymbol{p} = (1, 0, 0)$ is the only prior that causes this to happen. Nevertheless, this is sufficient for $\underline{P}_M(\{\boldsymbol{n}\}) = 0$. Importantly, the minimisation of posterior expectations under this likelihood over the set of priors M is not well-defined in the sense some priors cause the denominator in the Bayes' rule to be zero causing the objective function to be undefined at those points. On the other hand, the minimisation over $M - \{(1, 0, 0)\}$, as an infimum, is well-defined in this example.

■

There are two ways to proceed with our analysis. On the one hand, one can make posterior inference based on only the prior models that assign positive probability to the observed data. This is provided by the so-called *regular extension*.

Definition 6.2.1: (Walley [81], Appendix J.) For a set of distributions M, a suitably measurable set $B \subseteq \Omega$, the *regular extension (of $\underline{E}_M$ conditioning on B)* is defined as,

$$X \longmapsto \inf \left\{ \frac{E_P(I_B X)}{P(B)} : \ P \in M \ \wedge \ P(B) > 0 \right\}.$$

□

On the other hand, one can abandon the use of the generalised Bayes' rule and instead consider an alternative model. One established alternative method in this case is the so-called *vacuous extension*.

Definition 6.2.2: (Walley [81], 8.4.1) For a set of distributions M, a suitable measurable set $B \subseteq \Omega$, the *vacuous extension*[1] *(of $\underline{P}$ to conditioning on B)* is defined as,

$$X \longmapsto \left\{ \begin{array}{cc} \inf \left\{ \frac{E_P(I_B \cdot)}{P(B)} : \ P \in M \right\} & \text{if } \underline{P}_M(B) > 0, \\ \inf_{\omega \in \Omega} X(\omega) & \text{if } \underline{P}_M(B) = 0 \end{array} \right\}.$$

□

[1] In Walley [81], this is a natural extension of a conditional prevision due to the generalised Bayes' rule, but, to avoid confusion, we will call this the vacuous extension.

Notice that the *vacuous extension* results in effectively throwing away the information provided by M when the prior lower probability of the dataset is zero. Walley [81] notes that, under certain conditions, they are both jointly coherent with the prior lower expectation as discussed in Section 2.2.4. (In Section 6.3.3, we will see how certain assessments and elicitations may violate this.) As they stand, they are simply modelling choices one makes in anticipation of actually observing B whose lower probability is apriori zero.

Note that, when (the closure of) M contains at least one prior that assigns all of its mass to the set of θ's that assign zero likelihood to n_1, the lower probability of observing the data will always be zero. To identify these priors, write,

$$\mathcal{I}_{\boldsymbol{n}} = \{i \in \ 1, \ldots, |\Theta| : \ L(\boldsymbol{\theta}_i|\boldsymbol{n}) > 0\},$$

the indices of Θ whose likelihood values are strictly positive.

Proposition 6.2.1: If $\boldsymbol{n}$ is such that $L(\boldsymbol{n}|\theta_i) = 0$ for at least one $i = 1, \ldots, |\Theta|$, and there exists $\boldsymbol{p}_0$ in the closure of M such that $\sum_{i \notin \mathcal{I}_{\boldsymbol{n}}} p_0(\theta_i) = 1$, then $\underline{P}_M(\boldsymbol{n}) = 0$.

Proof: Notice that,

$$\underline{P}_M(\boldsymbol{n}) = \inf \left\{ \sum_{i=1}^{|\Theta|} p(\theta_i) L(\boldsymbol{n}|\theta_i) : \ p \in M \right\}.$$

It is clear that the choice of the limit point $\boldsymbol{p}_0$ will minimise this quantity to zero.

□

6.3 Posterior imprecise inference for discrete parameter and observation spaces

From Definition 6.2.2, the vacuous extension takes values from the GBR when the lower probability of the conditioning event is strictly positive and takes the infimum of the random variable otherwise. On the other hand, the computation of the regular extension is more involved, and we explore it under the discrete setting introduced in Section 6.1.

6.3.1 Computing the regular extension

Notice that the marginal probability,

$$\boldsymbol{L} \cdot \boldsymbol{p} = \sum_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\boldsymbol{n}) \boldsymbol{p}(\boldsymbol{\theta}),$$

is zero iff every term in the sum is zero individually. In turn, this is equivalent to the statement that for each $\boldsymbol{\theta} \in \Theta$, at least one of $L(\boldsymbol{\theta}|\boldsymbol{n})$ and/or $\boldsymbol{p}(\boldsymbol{\theta})$ are zero, and so the valid Bayes' rules come from priors that avoid this.

For constructing the regular extension, we identify the set of priors that yield strictly positive marginal probability for observing $\boldsymbol{n}$:

$$\{\boldsymbol{p} \in M : \ \exists\ i \in \mathcal{I}_{\boldsymbol{n}},\ p(\boldsymbol{\theta}_i) > 0\}.$$

The (lower and upper) regular extension(s) for some random variable $f(\boldsymbol{\theta})$ are,

$$\underline{R}_M(f(\boldsymbol{\theta})|\boldsymbol{n}) = \inf \left\{ \frac{\sum_{i \in \mathcal{I}_{\boldsymbol{n}}} f_i L_i p_i}{\sum_{i \in \mathcal{I}_{\boldsymbol{n}}} L_i p_i} \;\middle|\; \boldsymbol{p} \in M : \ (\exists\ i \in \mathcal{I}_{\boldsymbol{n}} : \ p_i > 0) \right\}. \tag{6.4}$$

$$\overline{R}_M(f(\boldsymbol{\theta})|\boldsymbol{n}) = \sup \left\{ \frac{\sum_{i \in \mathcal{I}_{\boldsymbol{n}}} f_i L_i p_i}{\sum_{i \in \mathcal{I}_{\boldsymbol{n}}} L_i p_i} \;\middle|\; \boldsymbol{p} \in M : \ (\exists\ i \in \mathcal{I}_{\boldsymbol{n}} : \ p_i > 0) \right\}. \tag{6.5}$$

6.3.2 Effects of likelihood on regular extension values

From the above construction, the regular extension only considers the optimisation over the posterior distributions normalised over only the prior probabilities of models $\theta \in \Theta$ which have a positive likelihood. In turn, the lower and upper regular extensions of, say $f(\theta)$, are bounded below and above only by the minimum and maximum of f taken over this active set of models. We can summarise this as follows.

Corollary 6.3.1: Consider the setting of Section 6.3.1. Recall that,

$$\mathcal{I}_{\boldsymbol{n}} = \{i \in \ 1, \ldots, |\Theta| : \ L(\boldsymbol{\theta}_i|\boldsymbol{n}) > 0\},$$

is the index set of observation models with a strictly positive likelihood value. Let M be any subset of the set of all prior distributions over Θ. For any bounded function $f(\theta)$ on

Θ, its regular extensions are bounded only by the values associated with strictly positive likelihood,

$$\min_{i \in \mathcal{I}_{\boldsymbol{n}}} f(\theta_i) \leq \underline{R}(f(\theta)|x) \leq \overline{R}(f(\theta)|x) \leq \max_{i \in \mathcal{I}_{\boldsymbol{n}}} f(\theta_i).$$

Proof: This follows immediately from the form of the lower and upper regular extensions in (6.4) and (6.5).

□

6.3.3 Numerical behaviour of posterior inference

Example 6.3.1: Continuing with Example 6.2.1, consider the case that out of $n = 10$ documents, the phrase of interest occur in $n_1 = 1$ of them. Let us plot the contours of the posterior expectations,

$$E(\theta|\boldsymbol{p} = (p_1, p_2, p_3),\ \boldsymbol{n} = (1, 9)) = \frac{0 \cdot p_1 + 0.1 \cdot 0.1 \cdot 0.9^9 p_2 + 0.25 \cdot 0.25 \cdot 0.75^9 p_3}{0 \cdot p_1 + 0.1 \cdot 0.9^9 p_2 + 0.25 \cdot 0.75^9 p_3},$$

as a function of p_1, p_2, p_3. By (6.3), the prior $(1, 0, 0)$ causes the marginal probability to be zero. Because the Bayes' rule of this prior is not defined under this dataset, the regular extension excludes this prior over M for $\boldsymbol{n} = (1, 9)$.

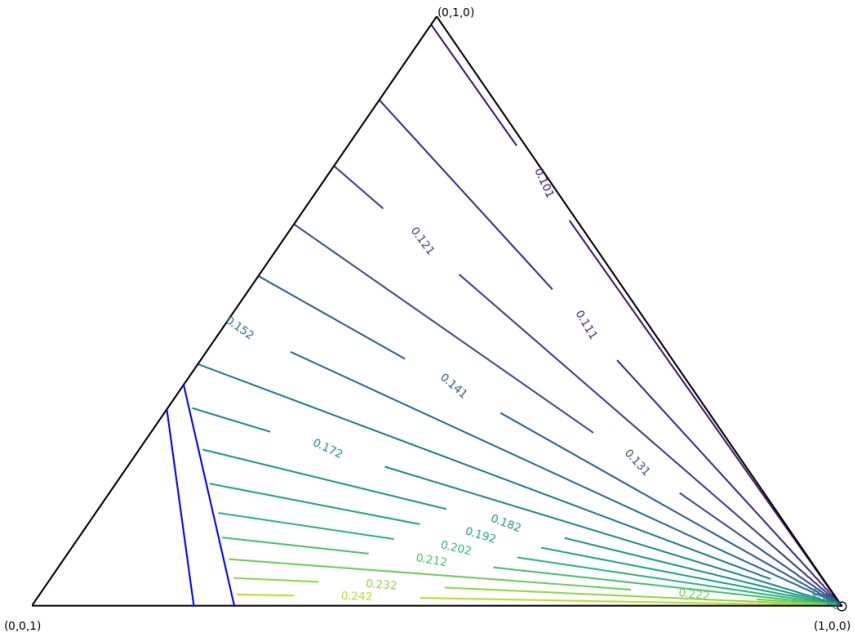

Figure 6.2: Contour map of the posterior expectation as a function of the projected prior distribution $(p_2, p_3) \longmapsto E(\theta|\boldsymbol{p} = (1 - p_2 - p_3, p_2, p_3),\ \boldsymbol{n} = (1, 9))$. Note that the $(1, 0, 0)$ is excluded from the polytope. Colours are normalised between 0.1 (dark purple) to 0.25 (light yellow).

Note that we cannot define the posterior expectation at this point of by appealing to continuity since the point is an intersection of the objective contours and so the limiting value is not well-defined. As a result, we have excluded this prior from the optimisation domain. With this, the minimum value of the regular extension is 0.1 and is achievable on the segment between $(0, 1, 0)$ and $(1, 0, 0)$ minus the latter endpoint. The maximum of 0.25 is achievable on the bottom segment between $(1/4, 0, 3/4)$ and $(1, 0, 0)$ minus the latter endpoint. Note that the posterior expectation of θ never achieves the minimum 0 over the set $\Theta = \{0, 0.1, 0.25\}$ in the regular extension's domain. On the other hand, the vacuous extension always yields $\inf_{\theta \in \Theta} \theta = 0$ as long as $(1, 0, 0)$ is in the convex region of the

optimisation.

The interpretation of whether or not 0.1 and 0.25 can be considered as posterior tight bounds reflecting the prior information depends on the rôle $(1, 0, 0)$ plays: in the case of the regular extension, the question is whether or not it can be ignored.

■

Example 6.3.2: We consider how the contour map changes when the dataset $\boldsymbol{n}$ is varied. Figure 6.3 displays the contour maps of the posterior expectation for various values of the sample size n and the observed proportion of documents in this corpus containing the phrase, $\widehat{\theta}_1$ (such that $n_1 = n\widehat{\theta}_1$), over values for $p_1.p_2, p_3$.

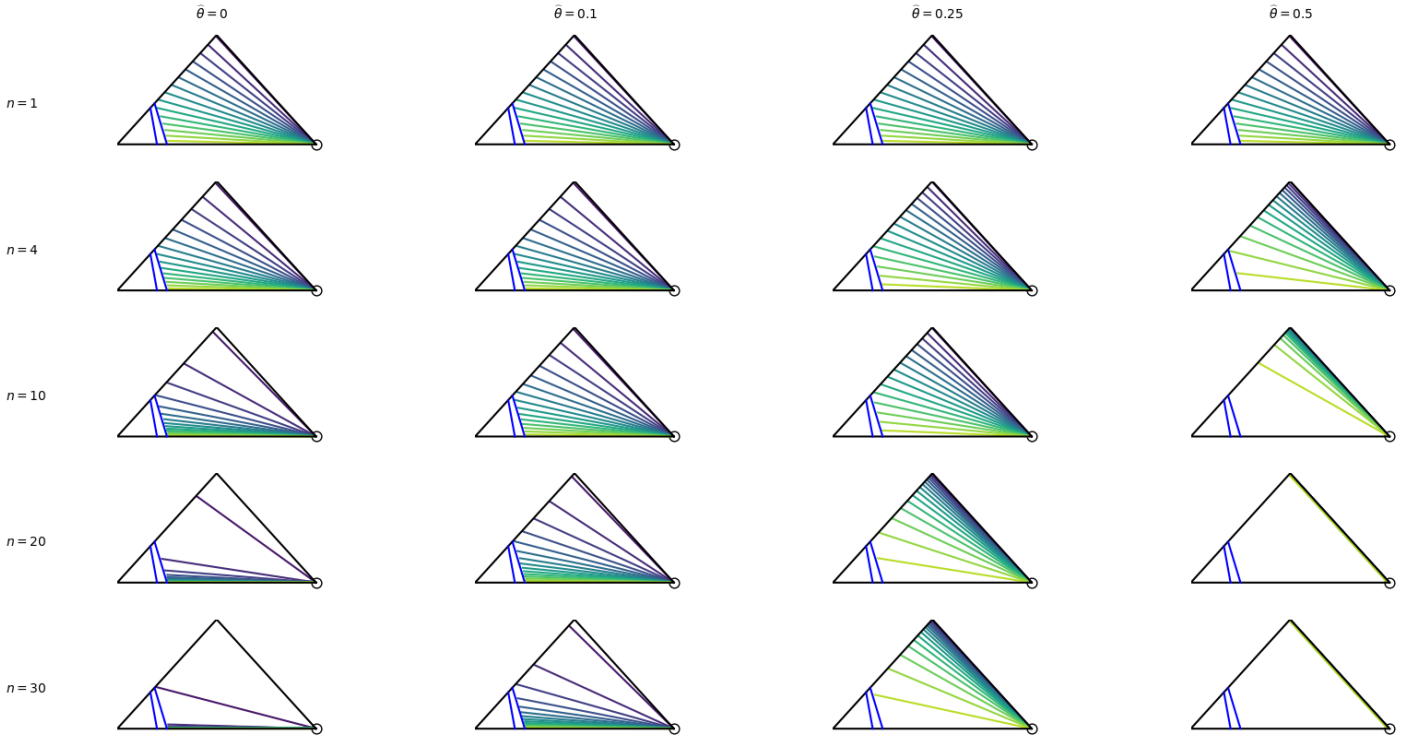

Figure 6.3: Contour maps of the posterior expectation as a function of the projected prior distribution $(p_2, p_3) \longmapsto E(\theta|\boldsymbol{p} = (1 - p_2 - p_3, p_2, p_3),\ \boldsymbol{n} = (n_1, n - n_1))$ for various values of n_1 and n. n is varied across the rows in $\{0, 4, 10, 20, 30\}$ and n_1 computed as $n\widehat{\theta}$ where $\widehat{\theta}$ is varied across the columns in $\{0, 0.1, 0.25, 0.5\}$. Colours are normalised between 0.1 (dark purple) to 0.25 (light yellow).

Figure 6.3 shows that the regular extension exhibits a typical imprecise behaviour as imprecise expectations as the dataset is varied. For example, as the number of observations n increases but the proportion of binomial successes remain zero (left column), the objective surface becomes flat and predominantly with values close to 0.1, the lower bound. Importantly, though, due to the likelihood effect described in Corollary 6.3.1, the lower bound remains strictly away from 0 even as no successes were observed.

■

Examples 6.3.1 and 6.3.2 demonstrate that, in our discrete context, the attainable values of the posterior expectation is restricted away from values for which the likelihood is zero. This is in accordance with Corollary 6.3.1. The following example explores the behaviour when the point removed by the regular extension $(1, 0, 0)$ is not included in the set of priors.

Example 6.3.3: Suppose that the linguist had elicited a third linear constraint on the possible prior distributions, resulting instead in the level curves of the posterior expectation in Figure 6.4 .

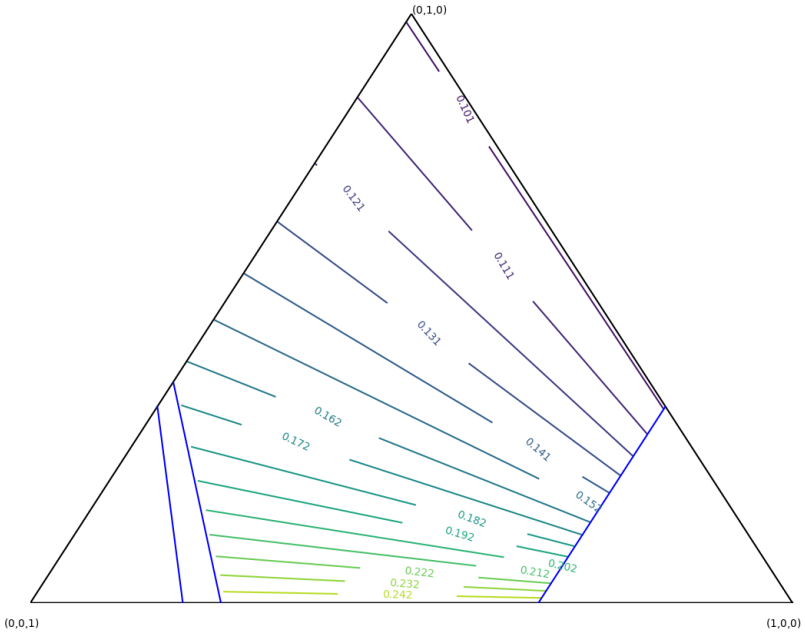

Figure 6.4: Contour map of the posterior expectation as a function of the projected prior distribution $(p_2, p_3) \longmapsto E(\theta|\boldsymbol{p} = (1 - p_2 - p_3, p_2, p_3),\ \boldsymbol{n} = (1, 9))$ with three constraints, one of which is redundant. Colours are normalised between 0.1 (dark purple) to 0.25 (light yellow).

We see that, the prior causing $\underline{P}(\{\boldsymbol{n}\}) = 0$, $(1, 0, 0)$ (by (6.3)), is no longer in the set, such that we can calculate the posterior lower expectation directly without either regular or vacuous extensions. Nevertheless, from the level curves, notice that the minimum value remains 0.1 which can be attained anywhere on the top-right edge of the convex set while the maximum remains 0.25 to be attained anywhere on the bottom edge.

■

In the following, we numerically explore the effects due to the assessment of the values of the likelihood on the regular extension. From Corollary 6.3.1 and the above examples, we have seen that the parameters for the observation model in Θ with a zero likelihood are ignored when evaluating the objective function of the regular extension. In the following examples, we produce two different assessments of the likelihood when zero success counts are observed: one assessment leads to a contradiction between the posterior regular extension and the prior imprecise model, while the other does not.

Example 6.3.4: Suppose that the linguist does not have enough information to impose *any* constraint on the set of priors, and $\boldsymbol{n} = (0, 10)$ is observed instead, resulting instead in the posterior expectation,

$$E(\theta|\boldsymbol{n} = (0, n), \boldsymbol{p}) = \frac{0 \cdot 0 \cdot p_1 + 0.1 \cdot 0.1^0(1 - 0.1)^{10}p_2 + 0.25 \cdot 0.25^0(1 - 0.25)^{10}p_3}{0 \cdot p_1 + 0.1^0(1 - 0.1)^{10}p_2 + 0.25^0(1 - 0.25)^{10} \cdot p_3},$$

over $\boldsymbol{p}$, whose the level curves are shown in Figure 6.5.

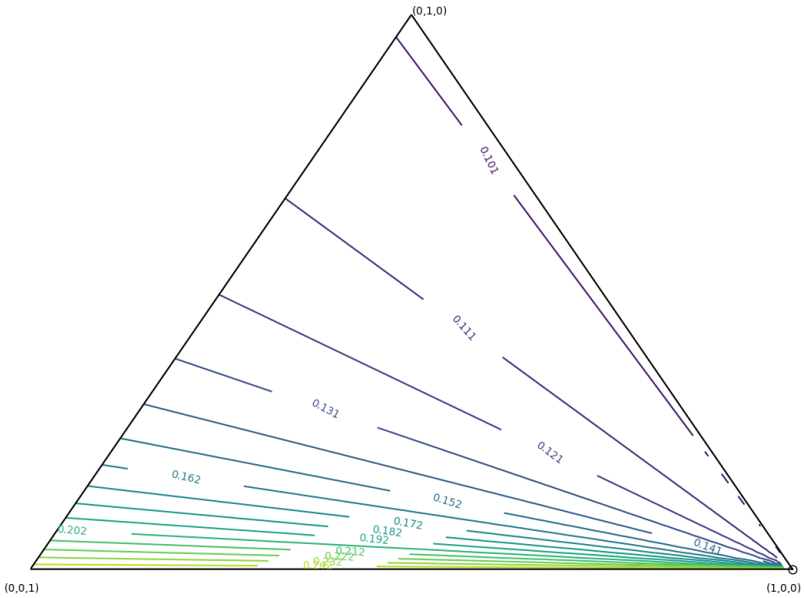

Figure 6.5: Contour map of the posterior expectation as a function of the projected prior distribution $(p_2, p_3) \longmapsto E(\theta|\boldsymbol{p} = (1 - p_2 - p_3, p_2, p_3),\ \boldsymbol{n} = (0, 10))$ with no constraints on the prior set except that the prior $(1, 0, 0)$ is excluded from the polytope. Colours are normalised between 0.1 (dark purple) to 0.25 (light yellow).

Starting with no constraint on the prior set and observing zero binomial successes, we see from the level curves that, again, the lower and upper regular extensions are attained at 0.1 on the right edge and 0.25 on the bottom edge, respectively. This example is of particular interest as it shows that the regular extension is inconsistent with the prior lower expectation in the sense that the lower regular extension that starts with no elicited constraints and observations with zero successes minimises to 0.1, such that the success probability $\theta = 0$ is ignored from Θ. This is in contrast with the prior lower expectation achieves the minimum of zero:

$$\underline{P}_M(\theta) = \min\ \{0 \cdot p_1 + 0.1 \cdot p_2 + 0.25 \cdot p_3 :\ (p_1, p_2, p_3) \in \overline{\triangle^3}\} = 0,$$

(minimised at $(1, 0, 0)$).

As in the examples in Section 6.3.3, this inconsistency between the regular extension and prior is due to the fact that, by Corollary 6.3.1, the value $\theta = 0$ is being ignored because the likelihood at $\theta = 0$ is zero under the dataset $\boldsymbol{n} = (0, 10)$. However, this also suggests that if the likelihood for $\theta = 0$ was *not* zero, then $\theta = 0$ can be exposed to the posterior expectation and this inconsistency can be avoided.

■

The issue in Example 6.3.4 pertains to the behaviour of the likelihood,

$$\theta^{n_1}(1-\theta)^{n-n_1}$$

around $\theta = 0$ and $n_1 = 0$. One way to specify a limit value of the likelihood at this point is to use the conventional definition that $0^0 = 1$ such that $\theta^{n_1}(1-\theta)^{n-n_1} = 1 \cdot 1^n = 1$. However, this limit is not unique: for example, 0 is also a limit if θ tends to 0 faster than n_1. Unlike the other examples where the dataset has at least one success (e.g. $\boldsymbol{n} = (1, 9)$) and the likelihood is well defined, we will be more explicit in assessing the value for $L(n_1 = 0|\theta = 0)$ in the following.

Example 6.3.5: Continuing with Example 6.3.4, let us directly assess a value for $L(n_1 = 0|\theta = 0)$ instead of relying on the multinomial likelihood's algebraic expression. Since $\theta = 0$ was originally elicited, we can interpret $L(n_1 = 0|\theta = 0) = P(n_1 = 0|\theta = 0)$ as the probability of observing no successes under a data generating process whose probability of success is zero: a reasonable assessment is $L(n_1 = 0|\theta = 0) = 1$. Using the rules of probability, the resulting posterior expectation is well defined for $\boldsymbol{n} = (0, 10)$ because the marginal probability is strictly positive:

$$E(\theta|\boldsymbol{n} = (0, 10), \boldsymbol{p}) = \frac{0 \cdot 1 \cdot p_1 + 0.1 \cdot 0.1^0(1-0.1)^{10}p_2 + 0.25 \cdot 0.25^0(1-0.25)^{10}p_3}{1 \cdot p_1 + 0.1^0(1-0.1)^{10}p_2 + 0.25^0(1-0.25)^{10} \cdot p_3}.$$

Notice that this conditional expectation is defined at $\boldsymbol{p} = (1, 0, 0)$ due to $L(n_1 = 0|\theta = 0) = 1$. Further, unlike in Example 6.3.4, this prior achieves the minimum posterior expectation of 0 over the set of priors M and coincides with the vacuous extension.

■

Examples 6.3.4 and 6.3.5 highlight the importance of the assessment of the likelihood function in contributing to the smooth operation of imprecise posterior inference in our cases.

In particular, the regular extension in posterior inferential settings is demonstrably sensitive to the definition (and any resulting pathologies) of the likelihood function.

6.4 Concluding remarks

Our contributions in this chapters are as follows. We demonstrated and compared the use of the regular and vacuous extensions in a concrete, finite dimensional setting with Examples 6.1.1, 6.2.1, 6.2.2, 6.3.1, 6.3.2 and 6.3.3 and remarked upon some characteristics of posterior inference when the set of priors is in conflict with data in the sense that the data's lower probability is zero. Corollary 6.3.1, Example 6.3.4 and Example 6.3.5 highlight the effect of the likelihood on the posterior regular extension and, importantly, how it may induce incoherence between the prior model and posterior regular extension in certain pathological cases.

The regular and vacuous extensions each have their use cases. The regular extension can be motivated as follows: if one observes a dataset, then its probability of being observed is, in hindsight, strictly positive such that it makes sense to perform posterior inference *as if* the assessed set of priors did not contain distributions assigning zero marginal probability to the observed dataset. Note that this does not violate the principle that the data should not influence the model choice because the choice of using the regular extension was apriori decided. The vacuous extension, on the other hand, effectively ignores all elicited prior information and produces vacuous bounds for the (conditional) expectations. It may sometimes be desirable to discard the entire prior elicitation. For example, if the set of priors were elicited using the same mechanism, then observing evidence against $P(B) = 0$ for certain P in the set may indicate that the elicitation mechanism that produced all the priors may be faulty, leading to the possibility that the elicitation for other P' with $P'(B) > 0$ being faulty as well.

The vacuous extension can be an overly conservative choice to handle conflict between prior and data. In cases such as Example 6.3.1, the only offending prior is a single vertex out of a polytope, and the geometry of the problem suggests that there are other points aside from the removed one that would have achieved the global minimum anyway. The vacuous extension would have missed the global minimum that effectively matches what we would have expected out of the sensitivity analysis over this polytope. Here, the vacuous extension is overly conservative on two counts. First, it ignores the information of

an entire set of elicited priors due to a single pathological one. Secondly, since $n_1 = 1$ in Example 6.3.1, it may be reasonable or even desirable to ignore the priors that assign a zero marginal probability to the dataset.

On the other hand, the regular extension can be overly optimistic (or precise). In corner cases such as Example 6.3.4, the behaviour of the posterior regular extension may be overly precise. In particular, the effect of the likelihood on posterior imprecise inference is interesting in the context of *zero-failure problems*. When the observation model space Θ apriori includes the possibility of a zero success probability and zero successes are later observed, coherence between the regular extension and the prior is dependent on the direct interpretation and assessment of the likelihood values where the expression of the likelihood function is ill-defined, as demonstrated in Examples 6.3.4 and 6.3.5.

Chapter 7

Geometry of conditioning on events with zero lower probability in finite dimensions

In Chapter 6, we have explored the choice between the vacuous and regular extensions for posterior inference when the lower marginal probability of the conditioning set is zero. Under this sort of conflict between the set of priors and the posterior resulting from conditioning upon such a dataset, we have observed that, in contrast with the posterior imprecise expectations over a set of prior distributions, the conditional vacuous extension effectively causes one to ignore one's unconditional set of priors while the regular extension continues by optimising only over the subset of priors that are not in conflict with the data. However, the principle of coherence by itself yields no compelling reason or definitive guideline on how to choose between the two in practice. Moreover, the analysis in Chapter 6 also has suggested that the regular and vacuous extensions may not be desirable as they are respectively overly precise or conservative in certain cases. This prompts us to consider possible models of intermediate levels of imprecision between the two extensions. In this chapter, we will consider the mathematical construction as well as the assessment of such extensions.

7.1 The existence of imprecise models between the vacuous and regular extensions

The existence of conditional assessments between the vacuous and regular extensions in the case when $\underline{P}_M(B) = 0$ is corroborated by Theorem 10 in Couso and Moral [29] in the literature of imprecise probabilities. We will not delve deeply into the result as it involves a significant departure from our current set of notations. Instead, we will summarise those concepts that are related to this chapter.

Couso and Moral investigate the conditioning problem using *sets of desirable gambles*, which are a set of random variables designated as a representation of a set of distributions in the space of random variables. The possible conditional models associated with a set of unconditional distributions are characterised by the choice of topological boundary points of the set of desirable gambles representing the unconditional model. In particular, the choice to exclude all boundary points results in the vacuous extension and there exists a maximal set of boundary points associated with the regular extension beyond which there is no gain in precision. Importantly, this suggests a class of extensions between this two associated with sets of boundary points which are non-empty strict subsets of the set associated with the regular extension.

7.2 Sets of conditional assessments between the vacuous and regular extensions

Let us consider a closed set of distributions M. Given a non-empty, suitably measurable subset $B \subseteq \Omega$, one can decompose M into,

$$M = \{P \in M : \ P(B) > 0\} \cup \{P \in M : \ P(B) = 0\}.$$

If $\{P \in M : \ P(B) = 0\}$ is empty, the generalised Bayes' rule (Proposition 2.2.1) yields a corresponding set of conditional distributions by applying Bayes' rule to each element of M. Otherwise, one way to form a conditional model is to use the regular extension introduced in Chapter 6, whose set of conditional distributions is,

$$M^{\mathrm{R}}_{|B} := \left\{ A \longmapsto \frac{P(AB)}{P(B)} : \ P \in M \ \wedge \ P(B) > 0 \right\}.$$

The unconditional distributions in $\{P \in M : \ P(B) = 0\}$ do not provide any information for constructing conditional distributions (see 6.10 of Walley [81]) and so cannot be used to construct probability assessments conditional on B. Recall that that the imprecision of a lower expectation decreases as the set of distributions over which one optimises for the lower envelope becomes smaller. In particular, the regular extension increases the precision of posterior inference by discarding distributions, and this may not accurately portray the degree of lack of information. One way to better reflect the state of knowledge conditional on B is to enlarge and append distributions to $M^{\mathrm{R}}_{|B}$.

Let us write $\mathcal{P}_B$ as the set of all distributions over the set of sample points B and $\underline{V}_B$ to be the vacuous lower expectation of $\mathcal{P}_B$. When $\underline{P}_M(B) = 0$, the vacuous extension corresponds exactly to

$$\underline{V}_B(X) = \inf_{\omega \in B} X(\omega),$$

where the vacuous extension on the right is achieved by a Dirac delta distribution in $\mathcal{P}_B$. Then,

$$M^{\mathrm{R}}_{|B} \subseteq \mathcal{P}_B,$$

such that we recover the fact that the vacuous extension is more imprecise than the regular extension:

$$\begin{aligned}
\underline{R}_M(X|B) &= \inf \ \{E_{P_B}(X) : \ P_B \in M^{\mathrm{R}}_{|B}\} \\
&\geq \inf \ \{E_{P_B}(X) : \ P_B \in \mathcal{P}_B\} \\
&= \inf_{\omega \in B} X(\omega) \\
&= \underline{V}_B(X).
\end{aligned}$$

This construction is suggestive of sets of distribution over B in the middle:

$$M^{\mathrm{R}}_{|B} \subseteq M^{\mathrm{R}}_{|B} \cup N \subseteq \mathcal{P}_B,$$

where N is any set of distributions over the sample points B[1]. It is clear that,

$$\underline{R}_M(X|B) \geq \inf \ \{E_{P_B}(X) : \ P_B \in M^{\mathrm{R}}_{|B} \cup N\} \geq \underline{V}_B(X).$$

[1]We implicitly assume that $M^{\mathrm{R}}_{|B} \cup N$ is closed so its lower expectation is attainable in the set.

7.2.1 Interpretation of N

We are interested in sets of distributions on B, N, that have the following methodological interpretation. Suppose that, for some fixed B, M was assessed such that $\underline{P}_M(B) = 0$. Then, B becomes observed, and this is in conflict with its lower probability being zero and requires a further investigation. Revisiting the information available apriori, the analysts come up with a set of distributions N over the sample points B in addition to $M^{\mathrm{R}}_{|B}$. We note that, therefore, N contains information only about random variables defined on B, not Ω, in replacement of such information not provided by unconditional distributions in M assigning a zero probability to B.

Example 7.2.1: Recall that, in Example 6.3.5, the indeterminacy of the limit of a binomial likelihood function $\theta^{n_1}(1-\theta)^{n-n_1}$ at $\theta = 0$ and $n_1 = 0$ caused the posterior expectation over M to be undefined at the prior $\boldsymbol{p} = (1, 0, 0)$ over the model space $\Theta = \{0, 0.1, 0.25\}$. As a result, the regular extension excludes this point from its domain of optimisation, M^{R}. Recall also that we have produced a posterior model alternative to the regular extension, by directly assessing the likelihood value $L(n_1 = 0|\theta = 0) = 1$ so that one defines, outside of the regular extension, the posterior distribution at $\boldsymbol{p} = (1, 0, 0)$ of θ as

$$\begin{aligned}
&\boldsymbol{p}_{|\boldsymbol{n}=(0,10),\boldsymbol{p}=(1,0,0)} \\
&\quad = \frac{1}{1 \cdot p_1 + 0.1^0(1-0.1)^{10}p_2 + 0.25^0(1-0.25)^{10} \cdot p_3} \begin{pmatrix} 1 \cdot p_1 \\ 0.1^0(1-0.1)^{10}p_2 \\ 0.25^0(1-0.25)^{10}p_3) \end{pmatrix} \\
&\quad = (1, 0, 0).
\end{aligned}$$

In other words, we have set $N = \{\boldsymbol{p}_{|\boldsymbol{n}=(0,10),\boldsymbol{p}}\}$ to be the posterior distribution in lieu of the ill-defined Bayes' rule posterior and the resulting lower expectation in Example 6.3.5 was in fact the optimisation over $M^{\mathrm{R}}_{|B} \cup \{\boldsymbol{p}_{|\boldsymbol{n}=(0,10),\boldsymbol{p}}\}$.

■

Even though one should in principle avoid revisiting the model after observing (a part of) the data, this situation is not completely unrealistic. As we have seen in Chapter 1, real-life elicitation processes are not exact and it is unreasonable to expect the resulting assessments to be exact. In practice, when a model M apriori assigns $\underline{P}_M(B) = 0$ and then B is observed, it is not unreasonable to revisit the elicitation process that has given rise to this assessment.

7.2.2 Intermediate extensions and joint coherence

Because we are effectively optimising over a larger set of distributions from $M^{\mathrm{R}}_{|B}$ to $M^{\mathrm{R}}_{|B} \cup N$, we can see that this new extension can be written as the minimum of the lower expectations of the components.

Definition 7.2.1: For $B \subseteq \Omega$, and N a set of distributions on B, define the lower expectation over $\mathcal{L}(\Omega)$,

$$\underline{E}^{\uparrow}_{N}(X|B) := \underline{E}_{N}(X|_{B}),$$

where $X|_B$ is the restriction of X to $B \subseteq \Omega$. For a partition $\mathcal{B}$ of Ω, write,

$$\underline{E}^{\uparrow}_{N}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} I_B \underline{E}^{\uparrow}_{N}(X|B).$$

□

Definition 7.2.2: For $B \subseteq \Omega$, the *intermediate extension of M by N conditional on B* is,

$$\underline{E}_{M;N}(X|B) := \min\left\{\underline{E}_{M^{\mathrm{R}}_{|B}}(X), \underline{E}^{\uparrow}_{N}(X)\right\},$$

For a partition $\mathcal{B}$ of Ω, the *intermediate extension of M by N conditional on $\mathcal{B}$* is,

$$\underline{E}_{M;N}(X|\mathcal{B}) := \sum_{B \in \mathcal{B}} I_B \underline{E}_{M;N}(X|B).$$

□

Its conjugate upper extension conditional on a set B,

$$\overline{E}_{M;N}(X|B) := -\underline{E}_{M;N}(-X|B),$$

is given by the maximum of its components:

$$\begin{aligned}
\overline{E}_{M;N}(X|B) &= -\min\left\{\underline{E}_{M^{\mathrm{R}}_{|B}}(-X), \underline{E}_{N}(-X|_{B})\right\} \\
&= -\min\left\{-\overline{E}_{M^{\mathrm{R}}_{|B}}(X), -\overline{E}_{N}(X|_{B})\right\} \\
&= \max\left\{\overline{E}_{M^{\mathrm{R}}_{|B}}(X), \overline{E}_{N}(X|_{B})\right\}.
\end{aligned}$$

Example 7.2.2: The vacuous extension is trivially recovered by setting N to be the space of all distributions over B, $\mathcal{P}_B$.

■

To preserve consistency between the unconditional and conditional assessments, they should be jointly coherent with each other. (See Section 2.2.4.) For example. when $\underline{P}_M(B) > 0$, the lower expectation due to generalised Bayes' rule in Proposition 2.2.1 is automatically jointly coherent with $\underline{E}_M$. The vacuous extension is jointly coherent with its prior model (8.4.1 of Walley [81]) and, under certain regularity conditions, the regular extension is also jointly coherent with its prior model. (See Appendix J. of Walley [81].)

We can show that with some reasonable assumptions, $\underline{E}_{M;N}(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$.

Theorem 7.2.1: (Theorem E.1.1) Let $\mathcal{B}$ be a partition of Ω, and M be a closed set of distributions. Suppose that N is such that $\underline{E}^{\uparrow}_N(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$. Suppose also that that such that its regular extension $\underline{R}_M(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$. Then, $\underline{E}_{M;N}(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$.

□

7.3 Elicitation and assessment of intermediate extensions

So far, we have defined the intermediate extensions in Definition 7.2.2. Furthermore, Theorem 7.2.1 establishes that the intermediate extension $\underline{E}_{M;N}(\cdot|\mathcal{B})$ is jointly coherent with the unconditional lower expectation $\underline{E}_M$. We now explore how the values of intermediate extensions can be assessed by incorporating certain types of additional information on how $\underline{P}_M(B) = 0$ can be further elicited.

7.3.1 Range of the intermediate extension

The fact that the lower intermediate extension is a minimum of the lower regular extension means that it is bounded above by the latter and, trivially, below by the lower vacuous extension. The conjugate upper intermediate extension $\overline{E}_{M;N}(\cdot|\mathcal{B})$ is similarly bounded above by the upper regular extension and, trivially, the upper vacuous extension. These all point to the fact that any lower bound of the conditional expectation that is above the regular extension is incoherent due to its being more precise than is justifiable by the information provided by M.

Example 7.3.1: Consider the extensions of a random variable X depicted in Figure 7.1.

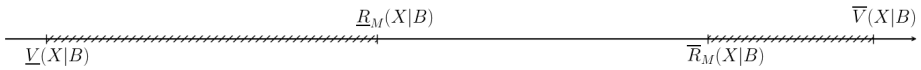

Figure 7.1: A schematic of regular extension $(\underline{R}_M(X|B), \overline{R}_M(X|B))$ and vacuous extension $(\underline{V}(X|B), \overline{V}(X|B))$ of X conditional on some event B. The values for any lower and upper intermediate extensions jointly coherent with the unconditional model $\underline{E}_M$ are hatched on the left and the right, respectively.

If one assesses a set of conditional distributions N such that its lower expectation $\underline{\mu}_N = \inf\{E_{Q_{|B}}(X|_B) : Q_{|B} \in N\}$ is in the left hatched area between $\underline{V}(X|B)$ and $\underline{R}_M(X|B)$, then, $\underline{E}_{M;N}(X|B) = \underline{\mu}_N$, since μ is less than the lower regular extension. Otherwise, $\underline{E}_{M;N}(X|B) = \underline{R}_M(X|B)$: $\underline{\mu}_N$ is overly precise and the regular extension 'corrects' it such that the intermediate extension is coherent.

■

Example 7.3.1 highlights an interpretation of N from an elicitation perspective. Because N always results in a non-decreasing change in precision, exogenous information that moves the lower intermediate extension strictly away from the lower regular extension indicates that there are conditional expectation values outside of $(\underline{R}_M(X|B), \overline{R}_M(X|B))$ that are consistent with M through joint coherence (when the information of N is also available). From an elicitation perspective, N may be treated as being representative of an expert

opinion that is *divergent* from the pool of experts responsible for eliciting M.

7.3.2 Examples of assessments

In this section, we illustrate the mechanism that determines the values of an intermediate expectation. We focus on finite dimensional cases where the denominator of Bayes' rule is zero for certain elements of a set of unconditional distributions. In such cases, one can assess conditional probabilities by further providing additional information in the form of the path that the denominator of Bayes' rule takes to approach zero.

Example 7.3.2: Consider a coherent unconditional set of distributions M over $\Omega = \{\omega_1, \omega_2, \omega_3\}$ such that $B = \{\omega_2, \omega_3\}$ and $\underline{P}_M(B) = 0$. When it is coherent, we know from Theorems 2.2.1 and 2.2.2 that $(1, 0, 0) \in M$ since it is the only distribution that achieves $\boldsymbol{p}(B) = 0$. This means that the point $(1, 0, 0)$ is an extreme point on the boundary of M. However, on observing B, the conditional expectation given by Bayes' rule is not well-defined at $(1, 0, 0)$: for example, in Figure 7.2, the level curves of the conditional expectation meet at the $(1, 0, 0)$. We can construct a (lower) intermediate extension by making a conditional assessment a limit of a set of values. We will explore the effects of the explicit choice of the limiting process on the intermediate extension conditional on B.

For simplicity, we assume that M is a convex polytope formed by a finite number of extreme points, of which $(1, 0, 0)$ achieves $\underline{P}_M(B) = 0$. Recall from Chapter 2 that every edge of M corresponds to an assessment that the expectation of a random variable Y is bounded below by (and attained at) some real number c, such that $\underline{P}_M(Y) = c$. This edge consists of the distributions $\{\boldsymbol{p} \in \overline{\triangle^3} : \ E_{\boldsymbol{p}}(Y) = c\}$. One can think of $\{(1, 0, 0)\}$ as the limit of a series of hyperplanes representing the moment condition $\underline{E}_{M_i}(Y_i) = c_i$ such that the hyperplane sequence eventually approaches and shrinks into the singleton set (see Figure 7.2).

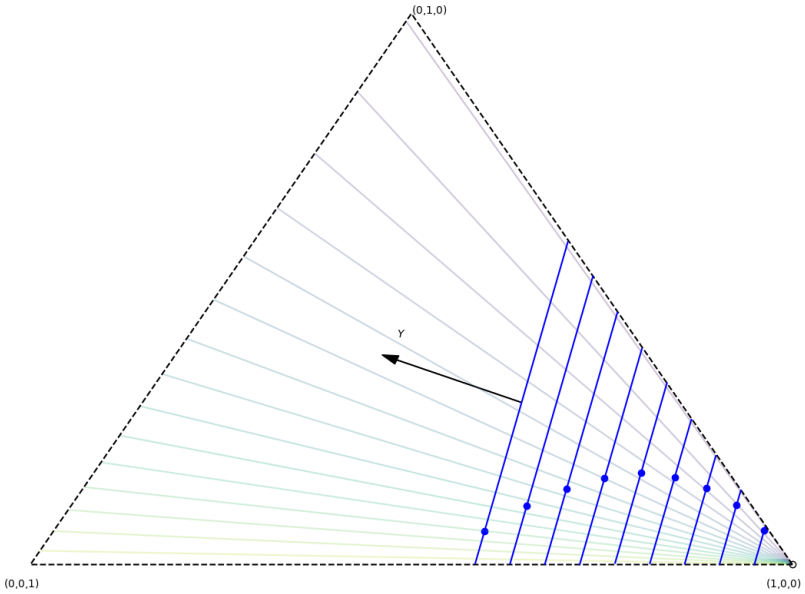

Figure 7.2: The optimisation domain of some posterior expectation as a function of priors over a $\Omega = \{\omega_1, \omega_2, \omega_3\}$. The level curves of this function are drawn with partial transparency. The right vertex $(1, 0, 0)$ is excluded as the level curves meet there such that the function is not well-defined at that point. A sequence of planes (blue lines) converging to the singleton set $\{(1, 0, 0)\}$ each with the same normal vector Y, a fixed random variable taking values (y_1, y_2, y_3) over Ω. Each plane is of the form $\{\boldsymbol{p} : \ y_1p_1 + y_2p_2 + y_3p_3 = c_i\}$ with Y being its normal vector and where $\{c_i\}$ is a sequence of intercepts that move the planes towards the limit as $i \to \infty$. The blue dots show one possible path $\{\boldsymbol{p}_i\}_{i=1}^{\infty}$ approaching $(1, 0, 0)$ with each $\boldsymbol{p}_i$ belonging to the i-th plane.

We can demonstrate the dependence of the set of conditional expectation of some X on the moment conditions that parametrise the sequence of hyperplanes approaching $(1, 0, 0)$. For simplicity, we will consider fixing a single random variable Y for the moment condition of interest. For every fixed i, the edge representing $\underline{E}_{M_i}(Y) = c_i$ are the distributions,

$$\{\boldsymbol{p} \in \overline{\triangle^3} : \ (1 - p_2 - p_3)y_1 + p_2y_2 + p_3y_3 = c_i\},$$

for a sequence of real numbers c_i. Suppose further that $\min(Y) = Y(\omega_1) = y_1$ and $y_3 > y_1$. As a result, because the distributions eventually converge to $(1, 0, 0)$, the unconditional mean will approach y_1. (No convex combination of y_2 and y_3 can result in y_1.) Therefore, it is reasonable to parametrise c_i as $y_1 + \epsilon_i$ where $\epsilon_i \to 0$.

We can study this convergence by analysing any arbitrary sequence of distributions $\{\boldsymbol{p}_i\}_{i=1}^{\infty}$ such that each $\boldsymbol{p}_i$ is in the i-th hyperplane. Then, we can parametrise $\boldsymbol{p}_i = (1 - p_{i2} - p_{i3}, p_{i2}, p_{i3})$ by p_{i2}, such that the moment condition can be written as,

$$y_1(1 - p_{i2} - p_{i3}) + y_2 p_{i2} + y_3 p_{i3} = y_1 + \epsilon_i.$$

This leads to,

$$p_{i3} = \frac{\epsilon_i}{y_3 - y_1} - p_{i2}\frac{y_2 - y_1}{y_3 - y_1}.$$

Now, consider the conditional expectation of the random variable of interest, X, on $B = \{\omega_2, \omega_3\}$ under such a distribution on this edge:

$$\frac{x_2 p_{i2} + x_3 p_{i3}}{p_{i2} + p_{i3}} = \frac{x_2 p_{i2} + x_3 \left(\frac{\epsilon_i}{y_3 - y_1} - p_{i2}\frac{y_2 - y_1}{y_3 - y_1}\right)}{p_{i2} + \left(\frac{\epsilon_i}{y_3 - y_1} - p_{i2}\frac{y_2 - y_1}{y_3 - y_1}\right)}.$$

For fixed choices of Y and X, this expression can take on various values depending on the paths taken for $p_{i2} \to 0$ and $\epsilon_i \to 0$. For example, suppose that $\epsilon_i = o(p_{i2})$ such that ϵ_i approaches zero faster than p_{i2}. The expression will approach the following convex combination of x_2 and x_3,

$$\frac{x_2 + x_3 \cdot \frac{y_2 - y_1}{y_3 - y_1}}{1 + \frac{y_2 - y_1}{y_3 - y_1}}.$$

If $y_2 = y_1$, then this expression becomes x_2. On the other hand, if $p_{i2} = o(\epsilon_i)$, then the expression approaches x_3. If we set $p_{i2} = \epsilon_i$, then, the expression approaches,

$$\frac{x_2 + x_3 \left(\frac{1}{y_3 - y_1} - \frac{y_2 - y_1}{y_3 - y_1}\right)}{1 + \left(\frac{1}{y_3 - y_1} - \frac{y_2 - y_1}{y_3 - y_1}\right)} = \frac{x_2 + x_3 \left(\frac{y_3 - y_2}{y_3 - y_1}\right)}{1 + \left(\frac{y_3 - y_2}{y_3 - y_1}\right)}.$$

In all, we have demonstrated that the posterior expectation can approach any value in the convex combination of x_2 and x_3 as M_i approaches M depending the Y that parametrises the edge that approaches $\{(1, 0, 0)\}$.

■

Example 7.3.3: Recall the running example of Chapter 6, Example 6.3.1. The linguist has elicited the model space $\Theta = \{0, 0.1, 0.25\}$ for θ, the probability that a randomly chosen document in a corpus contains a certain phrase. Suppose that a set of priors M containing the prior $\boldsymbol{p} = (1, 0, 0)$ was also elicited. Finally, suppose that it is observed that one document out of a sample of ten contains the phrase, such that $\boldsymbol{n} = (1, 9)$. The observation probabilities $L(\theta|\boldsymbol{n}) = \theta^1(1-\theta)^9$ for each probability in Θ are,

$$L(\boldsymbol{n}|\theta = 0) = 0, \quad L(\boldsymbol{n}|\theta = 0.1) = 0.1 \cdot 0.9^9, \quad \text{and } L(\boldsymbol{n}|\theta = 0.25) = 0.25 \cdot 0.75^9.$$

For $\boldsymbol{p}$ such that the marginal probability of $\boldsymbol{n}$ is positive, the posterior expectation of θ is then given by,

$$\frac{0 \cdot p_1 + 0.1 \cdot 0.1 \cdot 0.9^9 p_2 + 0.25 \cdot 0.25 \cdot 0.75^9 p_3}{0 \cdot p_1 + 0.1 \cdot 0.9^9 p_2 + 0.25 \cdot 0.75^9 p_3}.$$

Once again, note that this expression is not defined for $\boldsymbol{p} = (1, 0, 0)$ as the denominator is zero. This opens up the possibility of choosing either the vacuous extension or regular extension as we have done in Example 6.3.1, or constructing an intermediate extension by assessing the conditional expectation at $\boldsymbol{p} = (1, 0, 0)$ directly.

Let us consider an intermediate extension defined by the following limiting processes. Suppose that $k_2, k_3 > 0$ and a positive sequence $a_i > 0$ such that $0 \leq k_2 a_i + k_3 a_i \leq 1$ and $a_i \to 0$. Set

$$p_{i2} = k_2 a_i, \quad \text{and } p_{i3} = k_3 a_i.$$

Then, p_{i2} and p_{i3} converges to zero as $i \to \infty$ and the expression for the posterior expectation approaches,

$$\frac{0.1 \cdot 0.1 \cdot 0.9^9 k_2 + 0.25 \cdot 0.25 \cdot 0.75^9 k_3}{0.1 \cdot 0.9^9 k_2 + 0.25 \cdot 0.75^9 p_3 k_3}.$$

Because k_2 and k_3 represent information that is exogenous to the model, they can take any values, such that the posterior expectation can take any value between 0.1 and 0.25 depending on the values of k_2 and k_3.

■

7.3.3 Interpreting $\underline{P}_M(B) = 0$

It is important to note that the assessment $\underline{P}_M(B) = 0$ can result from an assessment of another random variable other than I_B. As in Example 7.3.2, $\underline{P}_M(B) = 0$ can be thought

of as the result of a limiting process of expanding the boundary associated with the moment condition on *any* random variable, e.g. $\{P:\ E_P(Y) = c\}$ such that the limit set M includes at least one distribution P_0 assigning $P_0(B) = 0$. In turn, this random variable specified a path for this limit at P_0 upon which the conditional assessments with P_0 also depend. Thus, the effect of the ill-definedness at P_0 propagates into the imprecise model M. From an elicitation perspective, if a limiting process provides a valid explanation for $\underline{P}_M(B) = 0$, it may be informative to understand from which part of the elicitation process this stems.

Example 7.3.4: One scenario that results in $\underline{P}_M(B) = 0$ is the resolution of the assessment. For example, if $B = \{\omega_2, \omega_3\}$ with $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and Y is a random variable such that $\min\{y_1, y_2, y_3\} = y_1$, the assessment that,

$$\underline{E}_M(Y) = y_1, \tag{7.1}$$

may result from approximating,

$$\underline{E}_{M_\epsilon}(Y) = y_1 + \epsilon, \tag{7.2}$$

with ϵ being a measurement error during the assessment process. As discussed in Example 7.3.2, no convex combination of y_2 and y_3 can result in y_1, so $\boldsymbol{p}_0 = (1, 0, 0)$ is the only distribution over all distributions over Ω that achieves $E_{\boldsymbol{p}_0}(Y) = y_1$ with $\boldsymbol{p}_0$ attaining the lower envelope of M when it is coherent.

However, as we have seen in Example 7.3.2, choosing a convergence rate for ϵ affects the limiting values of the conditional assessments. Furthermore, eliciting information on a different Y will also influence these limiting values. Methodologically, one should scrutinise how one arrives at the elicitation (7.2) and the assessment (7.1). Perhaps the equipment is not sensitive enough to measure ϵ at a sufficient resolution while measuring Y, so the experts report (7.1) instead of (7.2) due to rounding errors. Furthermore, perhaps a different Y' may be measured at a higher resolution and that $\underline{E}_{M'}(Y') = c$ may result in $\underline{P}_{M'}(B) > 0$. In other words, in considering the limitation of one's knowledge of Y, one may opt to discard this information, possibly in favour of another piece of information Y' that has been more reliably measured (or otherwise, more reliably known).

■

When the assessment leading to $\underline{P}_M(B) = 0$ is about B itself, it may be informative to understand what kind of elicitation the assessment $\underline{P}_M(B) = 0$ is ideally meant to represent in order to troubleshoot and correct the assessment that the lower probability of B

is zero. In particular, we briefly explore a hypothetical correction to weaken the strength of the assessment $\underline{P}_M(B) = 0$ after observing B. (It is hypothetical in the sense that it cannot be modelled by imprecise probabilities.)

Example 7.3.5: When $\underline{E}_M$ is coherent, the statement $\underline{P}_M(B) = 0$ is equivalent to the existence of a distribution $P_0 \in M$ that assigns a zero probability to the event B. That is to say,

$$\min \{P(B) : P \in M\} = 0.$$

Qualitatively, the analyst has effectively committed to the possibility that P_0 is a valid model. That is, the possibility that B is impossible is *definitively* consistent with M.

Contrast this with another possible (and *much weaker*) assessment that $\underline{P}_M(B) > 0$. Qualitatively, this means that the lowest (attainable) probability of B over M is *not* zero, but does not specify exactly what this value is. This more accurately describes the state of the modeller having apriori assigned $\underline{P}_M(B) = 0$ and later observed B: the modeller knows that B is now possible, but no single non-zero lower bound can be elicited.

In terms of convex sets of distributions, no single hyperplane boundary can be specified. Therefore, it is unclear that this assessment (or any other strict inequality bounds about expectations) can be explicitly expressed in the imprecise probability framework, and highlights a statistical elicitation issue that imprecision has difficulty in addressing.

■

7.4 Concluding remarks

The contributions of this chapter are as follows. We were motivated by Couso and Moral [29] to consider extensions as in Definition 7.2.2 to capture conditional imprecise assessments of the form of lower envelopes of distributions on B in the case when $\underline{P}_M(B) = 0$. Theorem 7.2.1 guarantees the joint coherence of the intermediate extensions when the exogenously appended set of distributions and regular extension from the original set are simultaneously jointly coherent with the unconditional model. These extensions are pointwise bounded between the vacuous and regular extensions: Example 7.3.1 illustrates the

intuition that intermediate extensions in the form of Definition 7.2.2 always respect the coherence of the two extremal extensions. Examples 7.3.2 and 7.3.3 illustrate the mechanism by which information exogenous to the information used to construct M can determine the value of the intermediate extension. In particular, we focussed on information in the form of the specification of a limiting process by which a sequence of M_i with $\underline{P}_{M_i}(B) > 0$ approaches M with $\underline{P}_M(B) = 0$. In Example 7.3.4 and Section 7.3.3, we touched upon some possible ways to elicit intermediate extensions by revisiting and further scrutinising the elicitation that led to $\underline{P}_M(B) = 0$ after B was observed.

The main difficulty in using intermediate extensions lies in their elicitation. Example 7.3.4 provides an example of an elicitation context which readily justifies the limiting process introduced in Example 7.3.2. It remains an open question as to how other forms of exogenous knowledge can be translated into similar assessments, or if there are other kinds of limiting process for $M_i \to M$. From a mathematical perspective, in higher dimensional spaces, the limit will generally be a face of the probability simplex and the limiting process will be required to assign a value for a conditional assessment to each of the points on this face. The mechanism in such cases remain to be explored, and the elicitation process to specify this process remains also unclear. Even more generally, the possibility of representing the exogenous information with a mathematical object other than limiting processes is an outstanding issue.

We also highlight the fact that, when the (lower) probability of the conditioning event is zero, then the intermediate extension is able to more flexibly accommodate exogenous information than a precise model with a single distribution. When assessing the intermediate extension, appending the exogenous information N to $M^{\mathrm{R}}_{|B}$ either decreases or does not change the precision of the posterior inference (as in Example 7.3.1). Because of this property, if the intermediate extension happens to take values from the regular extension, then information from M is still preserved. In contrast, if one uses a precise model involving a single prior distribution, one must necessarily discard the original model in favour of a different set of assessments that incorporate the new information.

Within the imprecise methodology, the elicitation process for the vacuous and regular extensions is different from that used for the intermediate extensions. The vacuous and regular extensions produce posterior inference based on an automatic rule for including and excluding models from the set of distributions M. The decision to use either the vacuous or regular extensions is simpler because it depends only on judgements upon the existing assessments M. In contrast, intermediate extensions are not automatic and require an

explicit intervention from the modeller to provide information N exogenous to the assessment process of M. One can *always* choose to use the vacuous and regular extensions but one cannot use (nontrivial) intermediate extensions if no additional information N is available. However, as we have remarked following Example 7.3.1, N can usefully represent new conditional beliefs which are divergent from those implied by M and Bayes' rule.

Chapter 8

Thesis summary

In this thesis, we began our enquiry by motivating the need for imprecise models in statistics through careful scrutiny of the elicitation process used to arrive at a probability model. In particular, we explored cases where a single probability distribution cannot represent the elicited information.

We focused on applying the imprecise probabilities framework developed in Walley [81] to problems frequently encountered in Statistics. In particular, *coherent* sensitivity analyses using imprecise probabilities were performed for inference regarding the posterior expectation and quantiles of log-odds statistics under the imprecise Dirichlet model (IDM) in Chapters 3 and 4. Chapter 5 explored the optimization aspects of the posterior expectation problem.

Chapters 6 and 7 explored the imprecise analogue of posterior inference when the marginal probability of the conditioning observed data is zero: that is, when the lower probability is zero over a set of distributions. In particular, Chapter 6 explored the use of the *regular* and *vacuous extensions* in the existing literature to construct coherent conditional assessments in this situation. Chapter 7 noted that, due to a result by Couso and Moral [29] these two extensions are respectively the least and most imprecise conditional models thay are jointly coherent with the prior imprecise model. This led to our construction of extensions that are of intermediate imprecision. These last two chapters represent our exploration into how posterior inference can be constructed under the imprecise methodology when the conditioning event has a lower probability of zero.

These topics represent our earnest attempt at bridging between imprecise probabilities and everyday statistical practices. This bridge is important because, as we have seen in Chapter 1, imprecise probabilities allows for a richer representation of elicited information

such that the statistical model can more naturally reflect the state of knowledge without unnecessarily restricting the model. For this reason, we hope that this thesis can bring the imprecise methodology to the needed attention of the statistical community at large.

References

[1] Agresti, Alan. "A survey of exact inference for contingency tables". In: *Statistical Science* 7.1 (1992), pp. 131–177.

[2] Agresti, Alan. *Analysis of Ordinal Categorical Data*. Second Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2010.

[3] Agresti, Alan. *Categorical Data Analysis*. Third Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2013.

[4] Alabdulhadi, Manal H, Coolen-Maturi, Tahani, and Coolen, Frank PA. "Nonparametric predictive inference for comparison of two diagnostic tests". In: *Communications in Statistics-Theory and Methods* (2020), pp. 1–17.

[5] Arrow, Kenneth J. *Social Choice and Individual Values*. Cowles Commission Monograph No. 12. John Wiley & Sons, New York; Chapman & Hall, London, 1951.

[6] Augustin, Thomas et al. *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2014.

[7] Benavoli, Alessio and Zaffalon, Marco. "A model of prior ignorance for inferences in the one-parameter exponential family". In: *Journal of Statistical Planning and Inference* 142.7 (2012), pp. 1960–1979.

[8] Benavoli, Alessio et al. "Imprecise Dirichlet process with application to the hypothesis test on the probability that X Y". In: *Journal of Statistical Theory and Practice* 9.3 (2015), pp. 658–684.

[9] Berger, James O. "An overview of robust Bayesian analysis". In: *TEST. An Official Journal of the Spanish Society of Statistics and Operations Research* 3.1 (1994), pp. 5–124.

[10] Berger, James O. "Robust Bayesian analysis: sensitivity to the prior". In: *Journal of Statistical Planning and Inference* 25.3 (1990), pp. 303–328.

[11] Berger, James O. *Statistical Decision Theory and Bayesian Analysis*. Second Edition. Springer Series in Statistics. Springer-Verlag, New York, 1985.

[12] Berger, James O. and Berliner, Mark L. "Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors". In: *The Annals of Statistics* 14.2 (1986), pp. 461–486.

[13] Bernard, Jean-Marc. "An introduction to the imprecise Dirichlet model for multinomial data". In: *International Journal of Approximate Reasoning* 39.2 (2005), pp. 123–150.

[14] Bickis, Miķelis. "Towards a geometry of imprecise inference". In: *International Journal of Approximate Reasoning* 83 (2017), pp. 281–297.

[15] Birch, M. W. "The detection of partial association. I. The 2×2 case". In: *Journal of the Royal Statistical Society. Series B. Methodological* 26 (1964), pp. 313–324.

[16] Birgin, Ernesto G., Martínez, José Mario, and Raydan, Marcos. "Nonmonotone spectral projected gradient methods on convex sets". In: *SIAM Journal on Optimization* 10.4 (2000), pp. 1196–1211.

[17] Borkar, V. S., Konda, V. R., and Mitter, S. K. "On de Finetti coherence and Kolmogorov probability". In: *Statistics & Probability Letters* 66.4 (2004), pp. 417–421.

[18] Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge university press, Cambridge, 2004.

[19] Cattaneo, Marco EGV and Wiencierz, Andrea. "Likelihood-based imprecise regression". In: *International Journal of Approximate Reasoning* 53.8 (2012), pp. 1137–1154.

[20] Chen, Junbin, Coolen, Frank PA, and Coolen-Maturi, Tahani. "On nonparametric predictive inference for asset and European option trading in the binomial tree model". In: *Journal of the Operational Research Society* 70.10 (2019), pp. 1678–1691.

[21] Clarke, Bertrand and Gustafson, Paul. "On the overall sensitivity of the posterior distribution to its inputs". In: *Journal of Statistical Planning and Inference* 71.1-2 (1998), pp. 137–150.

[22] Cooke, Roger M. *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Environmental Ethics and Science Policy Series. The Clarendon Press, Oxford University Press, New York, 1991.

[23] Coolen, Frank PA and Augustin, Thomas. "A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories". In: *International Journal of Approximate Reasoning* 50.2 (2009), pp. 217–230.

[24] Coolen-Maturi, Tahani, Coolen, Frank PA, and Alabdulhadi, Manal. "Nonparametric predictive inference for diagnostic test thresholds". In: *Communications in Statistics-Theory and Methods* 49.3 (2020), pp. 697–725.

[25] Coolen-Maturi, Tahani, Coolen, Frank PA, and Muhammad, Noryanti. "Predictive inference for bivariate data: Combining nonparametric predictive inference for marginals with an estimated copula". In: *Journal of Statistical Theory and Practice* 10.3 (2016), pp. 515–538.

[26] Couso, Inés, Álvarez-Caballero, Antonio, and Sánchez, Luciano. "Reconciling Bayesian and Frequentist Tests: the Imprecise Counterpart". In: *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*. PMLR. 2017, pp. 97–108.

[27] Couso, Inés and Dubois, Didier. "A general framework for maximizing likelihood under incomplete data". In: *International Journal of Approximate Reasoning* 93 (2018), pp. 238–260.

[28] Couso, Inés and Dubois, Didier. "An imprecise probability approach to joint extensions of stochastic and interval orderings". In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer. 2012, pp. 388–399.

[29] Couso, Inés and Moral, Serafín. "Sets of desirable gambles: conditioning, representation, and precise probabilities". In: *International Journal of Approximate Reasoning* 52.7 (2011), pp. 1034–1055.

[30] Couso, Inés, Moral, Serafín, and Sánchez, Luciano. "The behavioral meaning of the median". In: *Information Sciences. An International Journal* 294 (2015), pp. 127–138.

[31] Crisma, L, Gigante, P, and Millossovich, P. "A notion of coherent revision for arbitrary random quantities". In: *Journal of the Italian Statistical Society*. Statistical Methods and Applications 6.3 (1997), pp. 233–243.

[32] De Finetti, Bruno. "Foresight: Its logical laws, its subjective sources". In: *Breakthroughs in Statistics*. Springer, 1992, pp. 134–174.

[33] De Rainville, François-Michel et al. "DEAP: a python framework for evolutionary algorithms". In: *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*. GECCO '12. ACM, 2012, pp. 85–92.

[34] Dempster, Arthur P. "Upper and lower probabilities induced by a multivalued mapping". In: *The annals of mathematical statistics* (1967), pp. 325–339.

[35] Diaconis, Persi and Freedman, David. "On the consistency of Bayes estimates". In: *The Annals of Statistics* 14.1 (1986), pp. 1–26.

[36] Diaconis, Persi and Ylvisaker, Donald. "Conjugate priors for exponential families". In: *The Annals of Statistics* 7.2 (1979), pp. 269–281.

[37] Dong, Jianping and Simonoff, Jeffrey S. "A geometric combination estimator for d-dimensional ordinal sparse contingency tables". In: *The Annals of Statistics* 23.4 (1995), pp. 1143–1159.

[38] Fisher, Ronald A. *The Design of Experiments*. Second Edition. Oliver & Boyd, Edinburgh & London., 1937.

[39] Garthwaite, Paul H, Kadane, Joseph B, and O'Hagan, Anthony. "Statistical methods for eliciting probability distributions". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 680–701.

[40] Gastwirth, Joseph L. *Statistical Reasoning in Law and Public Policy: Tort Law, Evidence and Health*. Vol. 2. Academic Press, Boston, MA, 1988.

[41] Gelman, Andrew et al. *Bayesian Data Analysis*. Third Edition. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2014.

[42] Goodman, Jay H. *Existence of Compromises in Simple Group Decisions*. PhD. Thesis. ProQuest LLC, Ann Arbor, MI, 1988.

[43] Gustafson, P., Srinivasan, C., and Wasserman, L. "Local sensitivity analysis". In: *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994*. Oxford University Press, New York, 1996, pp. 197–210.

[44] Haldane, JBS. "The precision of observed values of small frequencies". In: *Biometrika* 35.3/4 (1948), pp. 297–300.

[45] Hill, Bruce M. "Parametric models for An: splitting processes and mixtures". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 55.2 (1993), pp. 423–433.

[46] Hill, Bruce M. "Posterior distribution of percentiles: Bayes' theorem for sampling from a population". In: *Journal of the American Statistical Association* 63.322 (1968), pp. 677–691.

[47] Holmes, R. B. *Geometric Functional Analysis and its Applications*. Vol. 24. Graduate Texts in Mathematics. Springer, 2012.

[48] Huber, Peter J. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1981.

[49] Jeffrey, Richard. *Subjective Probability: The Real Thing*. Cambridge University Press, 2004.

[50] Jeffreys, Harold. *Theory of Probability*. Third Edition. The International series of monographs on physics. Clarendon Press, Oxford, 1961.

[51] Kurtek, Sebastian and Bharath, Karthik. "Bayesian sensitivity analysis with the Fisher-Rao metric". In: *Biometrika* 102.3 (2015), pp. 601–616.

[52] Lindley, Dennis V. *Introduction to probability and statistics from a Bayesian viewpoint. Part I: Probability*. Cambridge University Press, New York, 1965.

[53] Lindley, Dennis V. *Understanding Uncertainty*. Revised Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2014.

[54] Lindsay, Bruce G. and Basak, Prasanta. "Moments determine the tail of a distribution (but not much else)". In: *The American Statistician* 54.4 (2000), pp. 248–251.

[55] Madansky, Albert. *Externally Bayesian Groups*. Tech. rep. RAND Corporation, Santa Monica, 1964.

[56] Magnus, Wilhelm, Oberhettinger, Fritz, and Soni, Raj Pal. *Formulas and Theorems for the Special Functions of Mathematical Physics*. Third Edition. Die Grundlehren der mathematischen Wissenschaften, Band 52. Springer-Verlag New York, New York, 1966.

[57] Maroufy, Vahed and Marriott, Paul. "Local and global robustness with conjugate and sparsity priors". In: *Statistica Sinica* 30 (2020), pp. 579–599.

[58] Marques, Filipe J, Coolen, Frank PA, and Coolen-Maturi, Tahani. "Introducing non-parametric predictive inference methods for reproducibility of likelihood ratio tests". In: *Journal of Statistical Theory and Practice* 13.1 (2019), pp. 1–14.

[59] Marriott, Paul. "On the local geometry of mixture models". In: *Biometrika* 89.1 (2002), pp. 77–93.

[60] McCulloch, Robert E. "Local model influence". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 473–478.

[61] Mohri, Mehryar and Roark, Brian. *Structural Zeros versus Sampling Zeros*. Tech. rep. Oregon Health & Science University, Portland, OR, 2005.

[62] O'Hagan, Anthony. *Research in Elicitation*. 2005. URL: http://www.tonyohagan.co.uk/academic/pdf/ElicRes.pdf.

[63] O'Hagan, Anthony et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, 2006.

[64] Perks, Wilfred. "Some observations on inverse probability including a new indifference rule". In: *Journal of the Institute of Actuaries* 73.2 (1947), pp. 285–334.

[65] Perolat, Julien et al. "Generalizing the Wilcoxon rank-sum test for interval data". In: *International Journal of Approximate Reasoning* 56 (2015), pp. 108–121.

[66] Quaeghebeur, Erik. "The CONEstrip algorithm". In: *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*. Springer, 2013, pp. 45–54.

[67] Rudin, Walter. *Principles of Mathematical Analysis*. Second Edition. McGraw-Hill Book Co., New York, 1964.

[68] Ruggeri, Fabrizio and Sivaganesan, Siva. "On a global sensitivity measure for Bayesian inference". In: *Sankhyā, The Indian Journal of Statistics, Series A* 62.1 (2000), pp. 110–127.

[69] Ruggeri, Fabrizio and Wasserman, Larry. "Infinitesimal sensitivity of posterior distributions". In: *Canadian Journal of Statistics* 21.2 (1993), pp. 195–203.

[70] Savage, Leonard J. *The Foundations of Statistics*. Revised Edition. Dover Publications, New York, 1972.

[71] Schervish, Mark J., Seidenfeld, Teddy, and Kadane, Joseph B. "The fundamental theorems of prevision and asset pricing". In: *International Journal of approximate Reasoning* 49.1 (2008), pp. 148–158.

[72] Seidenfeld, Teddy, Kadane, Joseph B., and Schervish, Mark J. "On the shared preferences of two Bayesian decision makers". In: *The Journal of Philosophy* 86.5 (1989), pp. 225–244.

[73] Shafer, Glenn et al. *A mathematical theory of evidence*. Vol. 1. Princeton university press Princeton, 1976.

[74] Smith, Cedric A. B. "Consistency in statistical inference and decision". In: *Journal of the Royal Statistical Society, Series B (Methodological)* 23.1 (1961), pp. 1–25.

[75] Troffaes, Matthias C. M. "Conditional lower previsions for unbounded random quantities". In: *Soft Methods for Integrated Uncertainty Modelling*. Springer, 2006, pp. 201–209.

[76] Troffaes, Matthias C. M. "Decision making under uncertainty using imprecise probabilities". In: *International Journal of Approximate Reasoning* 45.1 (2007), pp. 17–29.

[77] Troffaes, Matthias C. M. and de Cooman, Gert. "Extension of coherent lower previsions to unbounded random variables". In: *Intelligent Systems for Information Processing*. Elsevier, 2003, pp. 277–288.

[78] Troffaes, Matthias C. M. and de Cooman, Gert. *Lower Previsions*. John Wiley & Sons, 2014.

[79] Varadhan, Ravi and Gilbert, Paul. "BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function". In: *Journal of Statistical Software* 32.4 (2009).

[80] Walley, Peter. "Inferences from multinomial data: learning about a bag of marbles". In: *Journal of the Royal Statistical Society, Series B (Methodological)* 58.1 (1996), pp. 3–57.

[81] Walley, Peter. *Statistical Reasoning with Imprecise Probabilities*. Vol. 42. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1991.

[82] Walley, Peter, Pelessoni, Renato, and Vicig, Paolo. "Direct algorithms for checking consistency and making inferences from conditional probability assessments". In: *Journal of Statistical Planning and Inference* 126.1 (2004), pp. 119–151.

[83] Weichselberger, Kurt. "The theory of interval-probability as a unifying concept for uncertainty". In: *International Journal of Approximate Reasoning* 24.2 (2000), pp. 149–170.

[84] Williams, Peter M. "Coherence, strict coherence and zero probabilities". In: *Proceedings of the Fifth International Congress on Logic, Methodology and Philosophy of Science*. Vol. 6. 1975, pp. 29–33.

[85] Williams, Peter M. "Indeterminate probabilities". In: *Formal Methods in the Methodology of Empirical Sciences*. Springer, 1976, pp. 229–246.

[86] Williams, Peter M. "Notes on conditional previsions". In: *International Journal of Approximate Reasoning* 44.3 (2007), pp. 366–383.

[87] Yeh, James. *Real Analysis: Theory of Measure and Integration*. Second Edition. World Scientific Publishing Company, 2006.

[88] Zhu, Hongtu, Ibrahim, Joseph G., and Tang, Niansheng. "Bayesian influence analysis: a geometric approach". In: *Biometrika* 98.2 (2011), pp. 307–323.

APPENDICES

Appendix A

Appendix to Chapter 2

A.1 Geometrical interpretation of avoiding losses for probabilities

Following Borkar, Kinda and Mitter [17], we can ascribe a geometrical interpretation to the general case of avoiding all Dutch book arbitrage. We note that indeed, this corresponds to the geometrical analogy of the argument given by Theorem 2.5.5 of Walley [81], where avoiding Dutch book arbitrage is also known as *avoiding sure losses*.

We give the following geometrical intuition for the finite dimensional case.

Theorem A.1.1: Let P be a probability assessment over a finite number of events $A_1, \ldots, A_s \subseteq \Omega$, $s < \infty$, over a finite sample space Ω with $|\Omega| < \infty$. Then, P is such that

$$\forall c \in \mathbb{R}^s : \max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0,$$

iff,

$$(P(A_1), \ldots, P(A_s)) \in \mathrm{Conv}(\{(I_{A_1}(\omega), \ldots, I_{A_s}(\omega)) : \ \omega \in \Omega\}).$$

Proof: ($\Rightarrow$) Suppose that P is such that, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$,

$$\max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0.$$

After some manipulations, this is if and only if, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$, and $c_{s+1} \in \mathbb{R}$,

$$\max_{\omega \in \Omega} \left(\sum_{i=1}^{N} c^* I_{A_i}(\omega) + c_{s+1} \right) \geq \sum_{i=1}^{N} c_i P(A_i) + c_{s+1}.$$

This means that the following is false:

$$\exists \boldsymbol{c}^* \in \mathbb{R}^{s+1} : \left(\sum_{i=1}^{N} c_i^* I_{A_i}(\omega) + c_{s+1}^* \right) \leq 0 \ \wedge \ \sum_{i=1}^{N} c_i^* P(A_i) + c_{s+1}^* > 0.$$

But, notice that this can be put into the form,

$$\boldsymbol{A}^T \boldsymbol{c}^* \leq 0 \ \wedge \ \boldsymbol{b}^T \boldsymbol{c}^* > 0, \tag{A.1}$$

with $\boldsymbol{A}^T = [(I_{A_i}(\omega_j) : \ i = 1, \ldots, s, \ j = 1, \ldots, |\Omega|); \ \mathbf{1}]$ being the matrix formed by appending the one vector $\mathbf{1}$ to the right of the matrix $(I_{A_i}(\omega_j))_{ij}$ and $\boldsymbol{b}^T = (P(A_1), \ldots, P(A_s), 1)$. By Farkas' lemma, because the inequalities (A.5) have no solution, its dual

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b} \ \wedge \ \boldsymbol{c} \geq 0, \tag{A.2}$$

will have solution. But these conditions can be expanded as,

$$\begin{gathered} \forall j = 1, \ldots, |\Omega| : \ P(A_i) = \sum_{i=1}^{s} I_{A_i}(\omega_j) c_i, \\ \mathbf{1}^T \boldsymbol{c} = 1, \\ \boldsymbol{c} \geq 0. \end{gathered}$$

Equivalently, the vector $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$, as required.

($\Leftarrow$) Now suppose that $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$. Then, reversing Farkas' lemma implies (A.5), and retracing the equivalence of the statements above yield the converse.

□

A.2 Results and proofs

Lemma A.2.1: For any sample space Ω, and P a distribution over some σ-field of Ω, any expectation E_P over all bounded random variables avoids sure losses. In other words, for any $X_1, \ldots, X_n$ that are bounded,

$$\sup_{\omega} \sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) \geq 0.$$

Proof: Indeed, because E_P is an expectation and X_i's are bounded, the finite sum of expectations is the expectation of the sum, so that,

$$\sup_{\omega} \sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) = \sup_{\omega} \sum_{i=1}^{n} X_i - E_P\left(\sum_{i=1}^{n} X_i\right).$$

Because expectations are convex combinations of its components, we can write,

$$\sup_{\omega} \sum_{i=1}^{n} X_i \geq E_P\left(\sum_{i=1}^{n} X_i\right),$$

yielding the result.

□

Lemma A.2.2: For any sample space Ω, and P a distribution over Ω over some σ-field of Ω, any expectation E_P over all bounded random variables are coherent. In other words, for any $X_0, X_1, \ldots, X_n$ that are bounded and $m \in \mathbb{N}$,

$$\sup_{\omega} \sum_{i=1}^{n} ((X_i(\omega) - E_P(X_i)) - m(X_0 - E_P(X_0)) \geq 0.$$

Proof: Indeed, because E_P is an expectation and X_i's are bounded, the finite sum of expectations is the expectation of the sum, so that,

$$\sup_{\omega} \left(\sum_{i=1}^{n} (X_i(\omega) - E_P(X_i)) - m(X_0 - E_P(X_0))\right) = \sup_{\omega} \left(\sum_{i=1}^{n} X_i - mX_0\right) - E_P\left(\sum_{i=1}^{n} X_i - mX_0\right).$$

Because expectations are convex combinations of its components, it is clear that,

$$\sup_{\omega} \left(\sum_{i=1}^{n} X_i - mX_0 \right) \geq E_P \left(\sum_{i=1}^{n} X_i - mX_0 \right),$$

yielding the result.

□

Proposition A.2.1: Let $X = (-1/8, 1/4)$, $Y = (2, -1)$ and $Z = (3, -1/2)$. The assessments on p in the form,

$$E_P(X) \geq 0, \quad E_P(Y) \geq 0, \quad E_P(Z) \geq 0 \ ,$$

avoid sure losses over the domain $\mathcal{F} = \{X, Y, Z\}$.

Proof: Let us denote the lower bounds by,

$$\underline{E}(X) = 0, \ \underline{E}(Y) = 0, \underline{E}(Z) = 0.$$

We need to show the following,

$$\begin{aligned}
&\forall W \in \mathcal{F}: && \sup_{\omega} \left(W(\omega) - \underline{E}(W)\right) \geq 0, \\
&\forall W_1 \neq W_2 \in \mathcal{F}: && \sup_{\omega} \left(W_1(\omega) - \underline{E}(W_1)\right) + \left(W_2(\omega) - \underline{E}(W_2)\right) \geq 0, \\
&\forall W_1 \neq W_2 \neq W_3 \in \mathcal{F}: && \sup_{\omega} \left(W_1(\omega) - \underline{E}(W_1)\right) + \left(W_2(\omega) - \underline{E}(W_2)\right) + \left(W_3(\omega) - \underline{E}(W_3)\right) \geq 0.
\end{aligned}$$

Plugging in the assessments simplify the conditions.

$$\begin{aligned}
&\forall W \in \mathcal{F}: && \sup_{\omega} W(\omega) \geq 0, \\
&\forall W_1 \neq W_2 \in \mathcal{F}: && \sup_{\omega} \left(W_1(\omega)\right) + \left(W_2(\omega)\right) \geq 0, \\
&\forall W_1 \neq W_2 \neq W_3 \in \mathcal{F}: && \sup_{\omega} \left(W_1(\omega)\right) + \left(W_2(\omega)\right) + \left(W_3(\omega)\right) \geq 0.
\end{aligned}$$

Because every X, Y, Z has a positive component, the first set of conditions are satisfied. The second set of conditions is satisfied because,

$$X + Y = (15/8, -3/4), \quad X + Z = (23/8, -1/4), \quad Y + Z = (5, -3/2) \ ,$$

such that each sum has a positive component. Finally,

$$X + Y + Z = (39/8, -5/4),$$

so that the third condition is satisfied.

□

Theorem A.2.1: Let P be a probability assessment over a finite number of events $A_1, \ldots, A_s \subseteq \Omega$, $s < \infty$, over the finite sample space Ω with $|\Omega| < \infty$. Then, P is such that

$$\forall c \in \mathbb{R}^s : \max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0,$$

iff,

$$(P(A_1), \ldots, P(A_s)) \in \text{Conv}(\{(I_{A_1}(\omega), \ldots, I_{A_s}(\omega)) : \omega \in \Omega\}).$$

Proof: ($\Rightarrow$) Suppose that P is such that, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$,

$$\max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0.$$

After some manipulations, this is if and only if, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$, and $c_{s+1} \in \mathbb{R}$,

$$\max_{\omega \in \Omega} \left(\sum_{i=1}^{N} c^* I_{A_i}(\omega) + c_{s+1} \right) \geq \sum_{i=1}^{N} c_i P(A_i) + c_{s+1}.$$

This means that the following is false:

$$\exists \boldsymbol{c}^* \in \mathbb{R}^{s+1} : \left(\sum_{i=1}^{N} c_i^* I_{A_i}(\omega) + c_{s+1}^* \right) \leq 0 \ \wedge \ \sum_{i=1}^{N} c_i^* P(A_i) + c_{s+1}^* > 0.$$

But, notice that this can be put into the form,

$$\boldsymbol{A}^T \boldsymbol{c}^* \leq 0 \ \wedge \ \boldsymbol{b}^T \boldsymbol{c}^* > 0, \tag{A.3}$$

with $\boldsymbol{A}^T = [(I_{A_i}(\omega_j) : i = 1, \ldots, s, \ j = 1, \ldots, |\Omega|); \ \mathbf{1}]$ being the matrix formed by appending the one vector $\mathbf{1}$ to the right of the matrix $(I_{A_i}(\omega_j))_{ij}$ and $\boldsymbol{b}^T = (P(A_1), \ldots, P(A_s), 1)$. By Farkas' lemma, because the inequalities (A.5) have no solution, its dual

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b} \ \wedge \ \boldsymbol{c} \geq 0, \tag{A.4}$$

will have solution. But these conditions can be expanded as,

$$\forall j = 1, \ldots, |\Omega| : \ P(A_i) = \sum_{i=1}^{s} I_{A_i}(\omega_j) c_i,$$
$$\mathbf{1}^T \boldsymbol{c} = 1,$$
$$\boldsymbol{c} \geq 0.$$

Equivalently, the vector $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$, as required.

($\Leftarrow$) Now suppose that $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$. Then, reversing Farkas' lemma implies (A.5), and retracing the equivalence of the statements above yield the converse.

□

Theorem A.2.2: Let P be a probability assessment over a finite number of events $A_1, \ldots, A_s \subseteq \Omega$, $s < \infty$, over the finite sample space Ω with $|\Omega| < \infty$. Then, P is such that

$$\forall c \in \mathbb{R}^s : \ \max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0,$$

iff,

$$(P(A_1), \ldots, P(A_s)) \in \text{Conv}(\{(I_{A_1}(\omega), \ldots, I_{A_s}(\omega)) : \ \omega \in \Omega\}).$$

Proof: ($\Rightarrow$) Suppose that P is such that, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$,

$$\max_{\omega \in \Omega} \sum_{i=1}^{s} c_i (I_{A_i}(\omega) - P(A_i)) > 0.$$

After some manipulations, this is if and only if, for all $c_i \in \mathbb{R}$, $i = 1, \ldots, s$, and $c_{s+1} \in \mathbb{R}$,

$$\max_{\omega \in \Omega} \left(\sum_{i=1}^{N} c^* I_{A_i}(\omega) + c_{s+1} \right) \geq \sum_{i=1}^{N} c_i P(A_i) + c_{s+1}.$$

This means that the following is false:

$$\exists \boldsymbol{c}^* \in \mathbb{R}^{s+1} : \ \left(\sum_{i=1}^{N} c_i^* I_{A_i}(\omega) + c_{s+1}^* \right) \leq 0 \ \wedge \ \sum_{i=1}^{N} c_i^* P(A_i) + c_{s+1}^* > 0.$$

But, notice that this can be put into the form,

$$\boldsymbol{A}^T \boldsymbol{c}^* \leq 0 \ \wedge \ \boldsymbol{b}^T \boldsymbol{c}^* > 0, \tag{A.5}$$

with $\boldsymbol{A}^T = [(I_{A_i}(\omega_j) : \ i = 1, \ldots, s, \ j = 1, \ldots, |\Omega|); \ \mathbf{1}]$ being the matrix formed by appending the one vector $\mathbf{1}$ to the right of the matrix $(I_{A_i}(\omega_j))_{ij}$ and $\boldsymbol{b}^T = (P(A_1), \ldots, P(A_s), 1)$. By Farkas' lemma, because the inequalities (A.5) have no solution, its dual

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b} \ \wedge \ \boldsymbol{c} \geq 0, \tag{A.6}$$

will have solution. But these conditions can be expanded as,

$$\begin{gathered} \forall j = 1, \ldots, |\Omega| : \ P(A_i) = \sum_{i=1}^{s} I_{A_i}(\omega_j) c_i, \\ \mathbf{1}^T \boldsymbol{c} = 1, \\ \boldsymbol{c} \geq 0. \end{gathered}$$

Equivalently, the vector $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$, as required.

($\Leftarrow$) Now suppose that $(P(A_1), \ldots, P(A_s))$ is in the convex hill of the set of vectors $\{(I_{A_1}(\omega_j), \ldots, I_{A_s}(\omega_j)) : \ j = 1, \ldots, |\Omega|\}$. Then, reversing Farkas' lemma implies (A.5), and retracing the equivalence of the statements above yield the converse.

□

Appendix B

Appendix to Chapter 3

B.1 The lower expectation of the general log-odds statistic under the IDM

We are interested in the posterior inference of the following general log odds statistic of a categorical distribution θ over $m < \infty$ categories under the imprecise Dirichlet model (IDM),

$$T : \theta \longmapsto \log \frac{\prod_{A\in\mathcal{A}} \sum_{i\in A} \theta_i}{\prod_{B\in\mathcal{B}} \sum_{j\in B} \theta_j} = \sum_{A\in\mathcal{A}} \log \theta_A - \sum_{B\in\mathcal{B}} \log \theta_B,$$

where $\mathcal{A} = \{A_1, \ldots, A_r\}$ and $\mathcal{B} = \{B_1, \ldots, B_q\}$ are finite collections of events with $A_a, B_b \subseteq \{1, \ldots, m\}$. This statistic includes various ratios of probabilities of interest:

- $\mathcal{A} = \{A\}, \mathcal{B} = \{B\}$ yields the log odds ratio $\log \theta_A - \log \theta_B$,
- For events $C_1, \ldots, C_q$, $\mathcal{A} = \{C_1 \cap \ldots \cap C_q\}, \mathcal{B} = \{C_1, \ldots, C_q\}$ is a ratio that measures the validity of the independence statement,

$$\theta_{\cap_{C\in\{C_1,\ldots,C_q\}} C} = \prod_{C\in\{C_1,\ldots,C_q\}} \theta_C.$$

In this section, we assume that A and B are not non-empty, are not the set of all categories, and $A \neq B$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

The posterior lower expectation of the general log odds under the IDM is given by,

$$\underline{E}\left[\log\frac{\prod_{A\in\mathcal{A}}\sum_{i\in A}\theta_i}{\prod_{B\in\mathcal{B}}\sum_{j\in B}\theta_j}\,\middle|\,\boldsymbol{n},\nu\right]=\inf_{\boldsymbol{\alpha}\in\overline{\triangle^m}}\left\{P_{\text{Dir}}\left[\sum_{A\in\mathcal{A}}\log\theta_A\,\middle|\,\boldsymbol{n},\boldsymbol{\alpha},\nu\right]-P_{\text{Dir}}\left[\sum_{B\in\mathcal{B}}\log\theta_B\,\middle|\,\boldsymbol{n},\boldsymbol{\alpha},\nu\right]\right\}. \tag{B.1}$$

An expression for the posterior expectations is obtained as follows.

Proposition B.1.1: Let $\theta\sim\text{Dirichlet}(\nu\boldsymbol{\alpha}+\boldsymbol{n})$ with $m<\infty$. Then,

$$P\left[\log\left(\sum_{i\in A}\theta_i\right)\,\middle|\,\nu\boldsymbol{\alpha}\right]=\nu\left[\psi\left(\nu\sigma_A(\boldsymbol{\alpha})+\sigma_A(\boldsymbol{n})\right)-\psi\left(\nu\sigma_{\{1,\ldots,m\}}(\boldsymbol{\alpha})\right)\right],$$

where $\psi(x)=\frac{d}{dx}\log\Gamma(x)$ is the *digamma function*.

□

The digamma function and its derivative, the *trigamma function*, are plotted below,

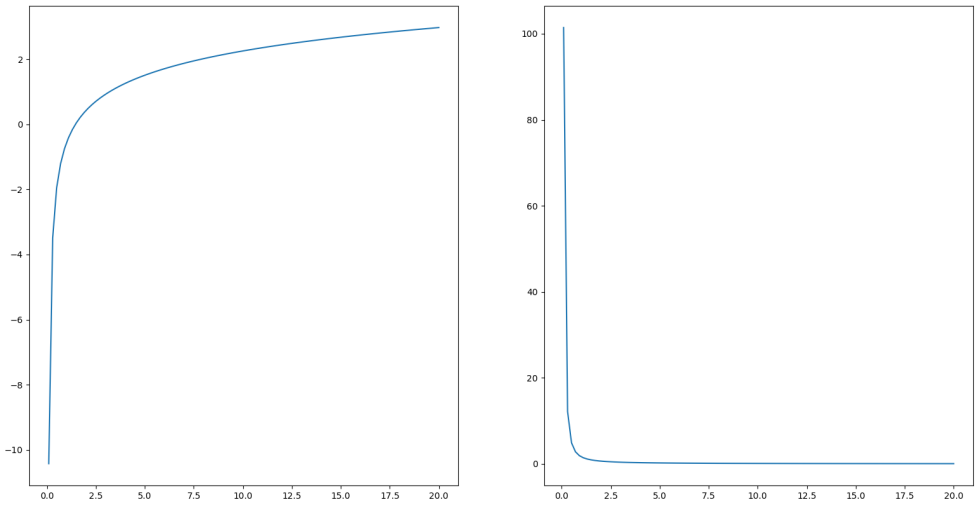

Figure B.1: The digamma (left) and trigamma (right) functins over $(0, 20]$.

We will be frequently invoking the facts that the diagamma function, ψ is an increasing function over $\mathbb{R}$, and that the trigamma function ψ' is a decreasing, strictly positive function.

Noting that the elements of $\boldsymbol{\alpha}$ sum to one such that, $\psi(\nu\sigma_{\{1,\ldots,p\}}(\boldsymbol{\alpha})) = \psi(\nu)$, the posterior lower expectation of the general log-odds under the IDM is,

$$\underline{E}\left[\log\frac{\prod_{A\in\mathcal{A}}\sum_{i\in A}\theta_i}{\prod_{B\in\mathcal{B}}\sum_{j\in B}\theta_j}\,\middle|\,\boldsymbol{n},\nu\right] = \min_{\boldsymbol{\alpha}\in\overline{\triangle^m}}\left\{\nu\left(\sum_{A\in\mathcal{A}}\psi(\nu\sigma_A(\boldsymbol{\alpha})) - \sum_{B\in\mathcal{B}}\psi(\nu\sigma_B(\boldsymbol{\alpha}))\right) - (|\mathcal{A}| - |\mathcal{B}|)\psi(\nu)\right\}. \tag{B.2}$$

Because the last term is independent of $\boldsymbol{\alpha}$, the achievable minimum is a solution to,

$$\begin{aligned}\text{minimise: } & \sum_{a=1}^{r}\psi(\nu\sigma_{A_a}(\boldsymbol{\alpha}) + \sigma_{A_a}(n)) - \sum_{b=1}^{q}\psi(\nu\sigma_{B_b}(\boldsymbol{\alpha}) + \sigma_{B_b}(n))\\ \text{subjected to: } & \boldsymbol{\alpha}\in\triangle^p.\end{aligned} \tag{B.3}$$

(It will be useful for what follows to index $\mathcal{A}$ by $a \in \{1,\ldots,r\}$ and $\mathcal{B}$ by $b \in \{1,\ldots,q\}$.)

B.2 Unboundedness of the log-odds and the theory of coherence

We detail the extension of Walley's [81] theory of coherence from bounded random variables for the IDM to the case of the unbounded log-odds random variable.

B.2.1 Extending Walley's [81] coherence to the log-odds random variable under the IDM under non-sparse observations

By a non-sparse observation, we mean that every category of the IDM system has at least one count being observed. The main result that drives our construction of the extension

is the convergence of a bounded approximation of the log-odds statistic.

Theorem B.2.1: Consider the general log-odds statistic,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

and the following truncation of it,

$$T_c(\boldsymbol{\theta}) = g(\boldsymbol{\theta})I(|g(\boldsymbol{\theta})| \leq c) + cI(|g(\boldsymbol{\theta})| > c).$$

Under $\boldsymbol{\theta} \sim \text{Dirichlet}(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ with sets $A_1, \ldots, A_r, B_1, \ldots, B_s$, such that $n_{A_i} > 0, n_{B_j} > 0$ for all A_i, B_j,

$$\sup_{\boldsymbol{\alpha} \in \overline{\triangle^{\Omega}}} E(|T_c - g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

Proof: see Theorem B.3.3.

□

Now, we define the IDM lower expectation of the unbounded log-odds statistic as a double limit of two limiting processes: the relaxation of the approximation of Theorem B.2.1 in the variable of c and the limit in the optimisation variable $\boldsymbol{\alpha}$ that may approach any point in $\overline{\triangle^m}$ (including the boundary).

We have, by the continuity of $\boldsymbol{\alpha} \longmapsto E(T_c|\nu\boldsymbol{\alpha} + \boldsymbol{n})$,

$$\lim_{c\to\infty} \left(\lim_{\boldsymbol{\alpha}\to\boldsymbol{\alpha}_0} E(T_c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) = \lim_{c\to\infty} E(T_c|\nu\boldsymbol{\alpha}_0 + \boldsymbol{n}) = E(g|\nu\boldsymbol{\alpha}_0 + \boldsymbol{n}).$$

On the other hand, by Theorem B.2.1,

$$\lim_{\boldsymbol{\alpha}\to\boldsymbol{\alpha}_0} \left(\lim_{c\to\infty} E(T_c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) = \lim_{\boldsymbol{\alpha}\to\boldsymbol{\alpha}_0} E(g|\nu\boldsymbol{\alpha}_0 + \boldsymbol{n}) = E(g|\nu\boldsymbol{\alpha}_0 + \boldsymbol{n}).$$

So, the double limits exist and are equal.

This justifies our plugging in of the log-odds into the IDM, despite it being unbounded, as follows. Any path based optimisation (such as gradient descent) over $\overline{\triangle^\Omega}$ of the expectation of the *truncated* log-odds T_c will yield converge to the limit point that is the infimum of the expectation of g, *as the truncation c is relaxed*. That is, we will make the following interpretation.

Definition B.2.1: For a coherent lower expectation, $\underline{E}$, defined over set of bounded random variables, write $\underline{E}^{(\text{ext})}$ to be its natural extension to the linear space of unbounded random variables.

□

Interpretation 2.2.1: For any general log-odds,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

whenever $\boldsymbol{n}$ contains at least one count in each category, we interpret,

$$\underline{E}^{(\text{ext})}_{\text{IDM}}(g(\boldsymbol{\theta})|\nu, \boldsymbol{n}) = \inf_{\boldsymbol{\alpha} \in \triangle^\Omega} E(g(\boldsymbol{\theta})|\nu\boldsymbol{\alpha} + \boldsymbol{n}),$$

as the double limit of the sequence of truncated Dirichlet expectations,

$$c, \boldsymbol{\alpha} \longmapsto E(g(\boldsymbol{\theta})I(|g(\boldsymbol{\theta})| \leq c) + cI(|g(\boldsymbol{\theta}) > c|)|\nu\boldsymbol{\alpha} + \boldsymbol{n}),$$

over the optimisation path of $\boldsymbol{\alpha}$ contained in $\overline{\triangle^\Omega}$ and the release of the truncation approximation via $c \to \infty$.

B.2.2 Convergence of lower and upper expectation of L^1 approximation error

We note an immediate corollary of Theorem B.2.1.

Corollary B.2.1: Under the conditions of Theorem B.2.1,

$$\inf_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

□

The uniform convergence in Corollary B.2.1 and Theorem B.2.1 are respectively equivalent to,

$$\underline{E}_{\text{IDM}}(|T_c - g|\ |\nu, \boldsymbol{n}) = \inf_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g|\ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

$$\overline{E}_{\text{IDM}}(|T_c - g|\ |\nu, \boldsymbol{n}) = \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g|\ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

B.2.3 Relation to coherence notions extended to unbounded random variables[78]

Theorem B.2.1 is closely related to the notion of previsible gambles introduced by Troffaes and de Cooman [78] defined in order to extend Walley's coherence to unbounded random variables. Roughly, a *previsible gamble* is one whose lower prevision can be approximated by a sequence of lower previsions of bounded gambles. We will show that any log-odds is previsible with respect to the posterior IDM previsions, and subsequently show that plugging in the general log-odds into the posterior IDM model is coherent in the sense of the extension of coherence to unbounded gambles by Troffaes and de Cooman.

Definition B.2.2: Throughout Section B.2.3 and nowhere else in this document, a *gamble* is a potentially unbounded random variable. This is in distinction to a *bounded gamble*. We will use these two terms explictly throughout Section B.2.3.

□

Definition B.2.3: (Troffaes and de Cooman [78], Definition 15.1, p.329) For a coherent lower prevision $\underline{E}$, and gambles f, f_n (which are not necessarily bounded), the net f_c *converges in* $\underline{E}$*-probability to* f iff,

$$\forall \epsilon > 0 : \ \overline{E}(|f - f_c| > \epsilon) \to 0,$$

(whereby we interpret 0 to be the Moore-Smith limit of $\overline{E}(|f - f_c| > \epsilon)$ as a net over c in an indexing directed set).

□

Definition B.2.4: (Troffaes and de Cooman [78], Definition 15.6, p.333) A gamble f is $\underline{E}$-*previsible* if there is a sequence f_n of bounded gambles such that,

1. f_n converges in $\underline{E}$-probability to f and,
2. $\lim_{n,m\to\infty} \overline{E}(|f_n - f_m|) = 0$.

If f is $\underline{E}$-previsible, then the *extended lower and upper previsions of* f are defined by,

$$\underline{E}^{\mathbf{x}}(f) := \lim_{n\to\infty} \underline{E}(f_n),$$

and

$$\overline{E}^{\mathbf{x}}(f) := \lim_{n\to\infty} \overline{E}(f_n),$$

respectively.

□

Proposition B.2.1: (Troffaes and de Cooman [78], Proposition 15.11, p 335) The previsions, $\underline{E}^{\mathbf{x}}$ and $\overline{E}^{\mathbf{x}}$ in Definition B.2.4 are coherent previsions over the set of $\underline{E}$-previsible gambles.

□

With these preliminaries from Troffaes and de Cooman [78], we can now show that the extension of the IDM to the unbounded log-odds gamble g is coherent by showing that it satisfies the conditions of Proposition B.2.1.

Theorem B.2.2: Under the conditions of Theorem B.2.1, the extended IDM whose value at the unbounded log-odds g is given by,

$$\underline{E}^{(\text{ext})}_{\text{IDM}}(g(\boldsymbol{\theta})|\nu, \boldsymbol{n}) = \lim_{c\to\infty} \underline{E}_{\text{IDM}}(T_c(\boldsymbol{\theta})|\nu\boldsymbol{\alpha} + \boldsymbol{n}),$$

is coherent under the Proposition B.2.1 of Troffaes and de Cooman [78].

Proof: By Proposition B.2.1, we are done if we show that, under the conditions of Theorem B.2.1, g is $\underline{E}_{\text{IDM}}(\cdot|\nu, \boldsymbol{n})$-previsible by showing that the two conditions in Definition B.2.4 is satisfied. These two conditions are shown to be satisfied by g (with the sequence of bounded gambles T_c) by Lemmas B.2.1 and B.2.2, respectively.

□

Lemma B.2.1: Under the conditions of Theorem B.2.1, the net T_c converges to g in $\underline{E}_{\text{IDM}}(\cdot|\nu, \boldsymbol{n})$-probability. That is,

$$\forall \epsilon > 0: \ \overline{E}_{\text{IDM}}(|T_c - g| > \epsilon|\nu, \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

Proof: Recall, by Theorem B.2.1,

$$\overline{E}_{\text{IDM}}(|T_c - g| \ |\nu, \boldsymbol{n}) = \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0.$$

On the other hand, by Markov's inequality,

$$\forall \epsilon > 0: \ P(|T_c - g| > \epsilon|\nu\boldsymbol{\alpha} + \boldsymbol{n}) < \frac{E(|T_c - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n})}{\epsilon},$$

such that,

$$\forall \epsilon > 0: \ \overline{E}_{\text{IDM}}(|T_c - g| > \epsilon|\nu, \boldsymbol{n}) = \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} P(|T_c - g| > \epsilon|\nu\boldsymbol{\alpha} + \boldsymbol{n}) < \frac{\sup_{\boldsymbol{\alpha} \in \overline{\triangle}} E(|T_c - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n})}{\epsilon} \to 0,$$

as $c \to \infty$.

□

Lemma B.2.2: Under the conditions of Theorem B.2.1,

$$\lim_{c_1, c_2 \to \infty} \overline{E}_{\text{IDM}}(|T_{c_1} - T_{c_2}| \ |\nu, \boldsymbol{n}) = 0.$$

Proof:

$$\begin{aligned}
&\overline{E}_{\text{IDM}}(|T_{c_1} - T_{c_2}| \ |\nu, \boldsymbol{n}) \\
&= \sup_{\boldsymbol{\alpha} \in \triangle^\Omega} E(|T_{c_1} - T_{c_2}| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \\
&\leq \sup_{\boldsymbol{\alpha} \in \triangle^\Omega} \left(E(|g - T_{c_1}| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) + E(|g - T_{c_2}| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n})\right) \\
&\leq \sup_{\boldsymbol{\alpha} \in \triangle^\Omega} E(|g - T_{c_1}| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) + \sup_{\boldsymbol{\alpha} \in \triangle^\Omega} E(|g - T_{c_2}| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \\
&\to 0. \qquad \text{(Theorem B.2.1)}
\end{aligned}$$

□

B.2.4 A note on behavioural interpretation of unbounded values of imprecise expectations [78]

Under the behavioural interpretation of Walley [81], $\underline{E}_{\text{IDM}}(f|\nu, \boldsymbol{n})$ is the highest price that the agent finds acceptable for the random reward f upon knowing only the IDM assumption, the fixed hyperparameter ν and the dataset $\boldsymbol{n}$. Specifically, $\underline{E}_{\text{IDM}}(f|\nu, \boldsymbol{n})$, is the price such that, for all $\epsilon > 0$, an agent knowing only this information is behaviourally disposed to engage in the random reward $f - \underline{E}_{\text{IDM}}(f|\nu, \boldsymbol{n}) + \epsilon$. (The meaning of 'behaviourally disposed' is extensively treated in Walley [81].)

Then, $\underline{E}_{\text{IDM}}(f|\nu, \boldsymbol{n}) = -\infty$ can be interpreted as: *there is no finite buying price below which a gambler who has this information at hand should find desirable.* (Similarly, $\overline{E}_{\text{IDM}}(f|\nu, \boldsymbol{n}) = -\infty$ can be interpreted as: *there is no finite selling price above which a gambler who has this information at hand should find desirable.*) This coincides exactly with the interpretation given by Troffaes and de Cooman for $\underline{E}(X) = -\infty$: for example, '*... [$\underline{E}(X) = -\infty$] is taken to mean that ... our subject is not willing to buy the gamble [X] for any real price $t \in \mathbb{R}$.*' and '*[$\underline{E}(X) = +\infty$] ... is taken to mean that ... the subject is willing to buy the gamble [X] at any price.*' [78]. In fact, the following remark by Troffaes and de Cooman is interesting to keep in mind.

> 'The rationality axioms [(that imply coherence)] are only an approximation of how agents really ought to behave... When utility becomes unbounded, we accept these rationality axioms as a matter of convenience in modelling... If things really matter at the infinite end, as with the Saint Petersburg paradox, perhaps one should rethink one's choice of utility.' [78]

B.2.5 A simple counterexample in the sparse data case

We can use the following counterexample to show that the limit of the approximation error fails to exist as we take both the relevant elements of the posterior hyperparameter, say $\boldsymbol{\gamma}$, to zero and the approximation parameter, c, tends to infinity.

Example B.2.1: Consider $\boldsymbol{\theta} \sim \text{Dirichlet}(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2) \in \triangle^{\Omega}$ and $\boldsymbol{n} = (n_1, n_2, n_3)$ and,

$$g(\boldsymbol{\theta}) = \log \theta_1/\theta_2,$$

for $\boldsymbol{\theta}$ a vector of trinomial cell probabilities. From Theorem B.2.1 for every fixed $\boldsymbol{\alpha} \in \triangle^{\Omega}$,

$$\lim_{c\to\infty} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = \lim_{c\to\infty} E(|g - gI(|g| \leq c)||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = 0,$$

This leads to,

$$\begin{gathered} \lim_{\alpha_1\to 0} \lim_{c\to\infty} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = 0 \\ \text{and } \lim_{\alpha_2\to 0} \lim_{c\to\infty} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = 0. \end{gathered} \quad . \tag{B.4}$$

However,

$$\begin{aligned} &\lim_{\alpha_1\to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) \\ &\geq \left(\lim_{\alpha_1\to 0} E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(|g|I(|g| \leq c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(cI(|g| > c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \\ &= \left(\lim_{\alpha_1\to 0} E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E(|g|I(|g| \leq c)|\nu\boldsymbol{\alpha} + \boldsymbol{n}) - cP(|g| > c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \\ &> \lim_{\alpha_1\to 0} E(|g||\nu\boldsymbol{\alpha} + \boldsymbol{n}) - 2c \\ &\geq \lim_{\alpha_1\to 0} |E(g|\nu\boldsymbol{\alpha} + \boldsymbol{n})| - 2c \\ &= \lim_{\alpha_1\to 0} |\psi^{(1)}(\nu\alpha_1 + n_1) - \psi^{(1)}(\nu\alpha_2 + n_2)| - 2c \qquad (\psi^{(1)} \text{ is the digamma function}) \\ &= \infty, \end{aligned}$$

and similarly, using the last two lines of the preceding inequalities,

$$\lim_{\alpha_2\to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) > \infty.$$

This leads to,

$$\begin{gathered} \lim_{c\to\infty} \lim_{\alpha_1\to 0} E(|g - T_c||\nu\boldsymbol{\alpha} + \boldsymbol{n}) = \infty \\ \text{and } \lim_{c\to\infty} \lim_{\alpha_2\to 0} E(|g - T_c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) = \infty \end{gathered} \quad . \tag{B.5}$$

This demonstrates that there are at least two paths in the (c, α_1, α_2) space that lead to different double limits of $E(|g - gI(|g| \leq c)||\nu\boldsymbol{\alpha} + \boldsymbol{n})$, such that the double limits are not defined.

■

B.3 Proof of Theorem B.3.1

For brevity, we will sometimes denote the general log odds by,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})}. \tag{B.6}$$

In this appendix section, we assume that, when dealing with log-odds involving any log probability of the form $\log P(A_i|\boldsymbol{\theta})$ where $P(A_i|\boldsymbol{\theta}) = \sum_{\omega \in A} \theta_\omega$ and $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\gamma})$, that $\gamma_A = \sum_{\omega \in A} \gamma_\omega > 0$. This is to ensure that the truncation is approximating something finite, or else convergence in mean and in probability are ill-defined.

B.3.1 Some lemmas

Lemma B.3.1: For a probability space over a finite dimensional simplex $(\overline{\triangle^\Omega}, \Sigma, \mu)$ such that $\mu(\triangle^\Omega) = 1$ and μ assigns zero probability to any lower dimensional subsets, and f which is $(\Sigma, \mathbb{B}_{\mathbb{R}\cup\{\pm\infty\}})$-measurable and μ-integrable, and c a natural number,

$$E|fI(-c \leq f \leq c) - f| \longrightarrow 0,$$

as $c \to \infty$.

Proof: Consider,

$$\{f_c : \ c \in \mathbb{N}\} = \{|fI(-c \leq f \leq c) - f| : \ c \in \mathbb{N}\}.$$

$\{f_c\}$ is a sequence of measurable functions in the index of c that decreases to zero in the interior of $\triangle^\Omega$, which has measure one under μ. Therefore, this convergence is μ-almost sure. For every fixed c, we have that,

$$f_c = |fI(-c \leq f \leq c) - f| = |-fI(|f| > c)| = |f|I(|f| > c) \leq |f|.$$

Then, by Corollary B.3.1, the relaxation of the monotone convergence theorem of Yeh [87] to almost sure convergence requirements, yields,

$$E|fI(-c \leq f \leq c) - f| \longrightarrow 0.$$

□

Lemma B.3.2: Under a Dirichlet distribution with parameters $\boldsymbol{\gamma} > \mathbf{0}$, the general log odds,

$$g(\boldsymbol{\theta}) = \log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})},$$

is integrable.

Proof: This follows from the existence of the second moments of $\log \theta_i$'s and cross moments $\log \theta_i \cdot \log \theta_j$, and Hölder's inequality:

$$E(|g||\boldsymbol{\gamma}) \leq (E(g^2|\boldsymbol{\gamma}))^{1/2} < \infty.$$

□

Lemma B.3.3: Under a Dirichlet distribution with parameters $\boldsymbol{\gamma} > \mathbf{0}$,

$$E \left| \left(\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})} \right) I \left(-c \leq \left(\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})} \right) \leq c \right) - \left(\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})} \right) \right| \longrightarrow 0.$$

Proof: follows from Lemma B.3.1 with $f = \left(\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})} \right)$ with the integrability of f guaranteed by Lemma B.3.2.

□

We will also need the following convergence in probability result.

Lemma B.3.4: Let $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\gamma})$ for some $\boldsymbol{\gamma} > \mathbf{0}$. Then,

$$cP \left(|\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})}| > c|\boldsymbol{\gamma} \right) \to 0,$$

and,

$$c^2 P \left(|\log \frac{\prod_{i=1}^{r} P(A_i|\boldsymbol{\theta})}{\prod_{j=1}^{s} P(B_j|\boldsymbol{\theta})}| > c|\boldsymbol{\gamma} \right) \to 0.$$

Proof: Write g for the general log odds as in (B.6). Consider that,

$$g^2 = \left(\sum_{i=1}^{r} \log \theta_{A_i} - \sum_{j=1}^{s} \log \theta_{B_j} \right)^2,$$

is an order two polynomial involving terms of the form $(\log \theta_A)^2$, $(\log \theta_A)(\log \theta_B)$, and the expectations of these terms all exist and are finite under the Dirichlet distribution. Importantly,

$$E\left(g^2|\boldsymbol{\gamma}\right) < \infty,$$

under the assumption that $\boldsymbol{\gamma} > 0$. So, using Markov's inequality,

$$cP\left(|g| > c|\boldsymbol{\gamma}\right) \leq c\frac{E\left(g^2|\boldsymbol{\gamma}\right)}{c^2} \to 0,$$

Similarly, the following fourth moment exists for the Dirichlet distribution when $\boldsymbol{\gamma} > 0$,

$$E\left(g^4|\boldsymbol{\gamma}\right) < \infty.$$

This yields,

$$c^2P\left(|g| > c|\boldsymbol{\gamma}\right) \leq c^2\frac{E\left(g^4|\boldsymbol{\gamma}\right)}{c^4} \to 0,$$

□

B.3.2 L^1 and pointwise convergence for general log-odds

Theorem B.3.1: (L^1 convergence) Under $\boldsymbol{\theta} \sim$ Dirichlet($\boldsymbol{\gamma}$) with sets $A_1, \ldots, A_r, B_1, \ldots, B_s$, such that $\gamma_{A_i} > 0, \gamma_{B_j} > 0$ for all A_i, B_j,

$$E(|T_c - g||\boldsymbol{\gamma}) \to 0.$$

as $c \to \infty$.

Proof: Lemma B.3.3 implies, as $c \to \infty$,

$$E(|g - gI(|g| \leq c)|\ |\boldsymbol{\gamma}) \to 0.$$

On the other hand, by Lemma B.3.4,

$$cP\left(|g| > c|\boldsymbol{\gamma}\right) \to 0,$$

as $c \to \infty$. In all,

$$E(|T_c - g|\ \boldsymbol{\gamma}) = E(|gI(|g| \leq c) - g + cI(|g| > c)|\ |\boldsymbol{\gamma}) \leq E(|gI(|g| \leq c) - g|\ |\boldsymbol{\gamma}) + cP(|g| > c|\boldsymbol{\gamma}) \to 0.$$

□

Theorem B.3.2: (Pointwise convergence) Under $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\gamma})$ with sets $A_1, \ldots, A_r, B_1, \ldots, B_s$, such that $\gamma_{A_i} > 0, \gamma_{B_j} > 0$ for all A_i, B_j,

$$E(T_c|\boldsymbol{\gamma}) \to E\left(g|\boldsymbol{\gamma}\right).$$

as $c \to \infty$.

Proof: this is a corollary of the L^1 convergence in Theorem B.3.1.

□

B.3.3 Uniform L^1 convergence for Dirichlet-Multinomial posterior expectations of general log-odds under non-sparse case

From the last section, we had the following pointwise convergence result,

$$E(|T_c - g||\boldsymbol{\gamma}) \to 0,$$

as $c \to \infty$, under the assumption that $\boldsymbol{\gamma}$ is such that $\gamma_{A_i}, \gamma_{B_j} > 0$ for all the sets A_i, B_j involved in the general log odds.

We remark that it is generally impossible for this continuity to be uniform when certain elements of $\boldsymbol{\gamma}$ are zero. For example, for a fixed general log odds involving sets A_1 and B_1, if $\gamma_{A_1}, \gamma_{B_1} = 0$, then the expectation of the general log odds becomes undefined as the difference of digamma functions $\psi(\gamma_{A_1}) - \psi(\gamma_{B_1})$ has ill-defined limiting behaviour. Rather, we will restrict ourselves to the special case of,

$$\boldsymbol{\gamma} = \nu\boldsymbol{\alpha} + \boldsymbol{n},$$

such that $\boldsymbol{n} > \mathbf{0}$ (elementwise), $\nu > 0$ and $\boldsymbol{\alpha} \in \overline{\triangle^{\Omega}}$. In this notation, the pointwise convergence is,

$$E(T_c|\nu\boldsymbol{\alpha} + \boldsymbol{n}) - E\left(g|\nu\boldsymbol{\alpha} + \boldsymbol{n}\right) \to 0.$$

Uniformity of convergence over the compact domain $\overline{\triangle^{\Omega}} \ni \boldsymbol{\alpha}$ is now much easier to obtain than over the positive orthant containing $\boldsymbol{\gamma}$ whose elements can be zero. Specifically, we

will verify the conditions of Dini's theorem, Theorem B.3.5, and apply it to obtain uniform convergence. We first present the proof of the main results, followed by the major lemmas used in it.

Theorem B.3.3: (Uniform L^1 convergence of posterior expectations) Under $\boldsymbol{\theta} \sim \text{Dirichlet}(\nu\boldsymbol{\alpha} + \boldsymbol{n})$ with sets $A_1, \ldots, A_r, B_1, \ldots, B_s$, such that $n_{A_i} > 0, n_{B_j} > 0$ for all A_i, B_j,

$$\sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g| | \nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

as $c \to \infty$.

Proof: First. consider that,

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|T_c - g| \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}) \\
&\leq \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|gI(|g| \leq c) - g| \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}) + c \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} P(|g| \geq c \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}) \\
&\leq \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(|gI(|g| \leq c) - g| \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}) + \frac{c}{c^2} \sup_{\boldsymbol{\alpha} \in \overline{\triangle^\Omega}} E(g^2 \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}). \quad \text{(Markov's Inequality)}
\end{aligned}$$

Because $\boldsymbol{n} > \boldsymbol{0}$, the last quantity is finite and so tends to zero in the limit of c. So, if we can show that the first supremum also tends to zero as $c \to \infty$, then we are finished.

We can use Dini's theorem to do so as follows. Let us verify the conditions for Dini's theorem. Over the compact domain $\overline{\triangle^\Omega} \ni \boldsymbol{\alpha}$, let

$$f_c(\boldsymbol{\alpha}) = E(|gI(|g| \leq c) - g| \ | \nu\boldsymbol{\alpha} + \boldsymbol{n}),$$

and

$$f(\boldsymbol{\alpha}) = 0.$$

- By Lemma B.3.5, f_c is continuous over $\overline{\triangle^\Omega}$ metrised by an L^p norm for every $c \in \mathbb{N}$.
- For every fixed $\boldsymbol{\alpha} \in \overline{\triangle^\Omega}$, $f_c \to f$, by the L^1 convergence in Lemma B.3.3.
- Since $|gI(|g| \leq c) - g| \geq |gI(|g| \leq c+1) - g|$, taking expectations on both sides for every fixed $\boldsymbol{\alpha} \in \overline{\triangle^\Omega}$, $f_c \geq f_{c+1}$.

Then, by Dini's theorem, the convergence,

$$E(|gI(|g| \leq c) - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}) \to 0,$$

is uniform over $\overline{\triangle^{\Omega}} \ni \boldsymbol{\alpha}$.

□

Lemma B.3.5: For every fixed $c \in \mathbb{N}$,

$$f_c(\boldsymbol{\alpha}) = E(|gI(|g| \leq c) - g| \ |\nu\boldsymbol{\alpha} + \boldsymbol{n}),$$

is continuous at every point in $\overline{\triangle^{\Omega}} \ni \boldsymbol{\alpha}$ where $\overline{\triangle^{\Omega}}$ is metrised by an L^p norm.

Proof: For short in this proof, write,

$$p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\nu\boldsymbol{\alpha} + \boldsymbol{n}).$$

Let us work in the projected parametrisation: write,

$$\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_{|\Omega|-1}) \in \overline{\blacktriangle^{|\Omega|-1}},$$

where

$$\overline{\blacktriangle^{|\Omega|-1}} := \left\{ \boldsymbol{p} \in \mathbb{R}^{|\Omega|-1} : \ p_i \geq 0 \ \wedge \ \sum_{i=1}^{|\Omega|-1} p_i \leq 1 \right\}.$$

such that,

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}'; \alpha_{|\Omega|}].$$

Consider that, for any two points $\boldsymbol{\alpha}', \boldsymbol{\alpha}'_0 \in \overline{\blacktriangle^{|\Omega|-1}}$,

$$\begin{aligned} f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0) &= \int_{\triangle^{\Omega}} |gI(|g| \leq c) - g|(\boldsymbol{\theta})(p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}) - p_{\boldsymbol{\alpha}'_0}(\boldsymbol{\theta}))d\boldsymbol{\theta} \\ &= \int_{\triangle^{\Omega}} |g|I(|g| > c)(\boldsymbol{\theta})(p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}) - p_{\boldsymbol{\alpha}'_0}(\boldsymbol{\theta}))d\boldsymbol{\theta}. \end{aligned}$$

It now suffices to prove continuity over $\overline{\blacktriangle^{|\Omega|-1}}$ to obtain continuity over $\overline{\triangle^{\Omega}}$.

Because $\boldsymbol{n} > 0$, for every fixed $\boldsymbol{\theta} \in \triangle^{\Omega}$, the gradient of the map $\boldsymbol{\alpha}' \longmapsto p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta})$ has components,

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_i} p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}) &= \frac{\partial}{\partial \alpha_i} \left(\frac{1}{B(\nu\boldsymbol{\alpha} + \boldsymbol{n})} \prod_{j=1}^{|\Omega|-1} \theta_j^{\nu\alpha_j + n_j - 1} \theta_{|\Omega|}^{\nu(1-\alpha_1-\ldots-\alpha_{|\Omega|-1})+n_{|\Omega|}-1} \right) \\
&= \frac{1}{B(\nu\boldsymbol{\alpha} + \boldsymbol{n})^2} \left(B(\nu\boldsymbol{\alpha} + \boldsymbol{n}) \left(\frac{\partial}{\partial \alpha_i} \prod_{j=1}^{|\Omega|-1} \theta_j^{\nu\alpha_j + n_j - 1} \theta_{|\Omega|}^{\nu(1-\alpha_1-\ldots-\alpha_{|\Omega|-1})+n_{|\Omega|}-1} \right) \right. \\
&\qquad \left. - \prod_{j=1}^{|\Omega|} \theta_j^{\nu\alpha_j + n_j - 1} \left(\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \right) \\
&= \frac{1}{B(\nu\boldsymbol{\alpha} + \boldsymbol{n})^2} \left(B(\nu\boldsymbol{\alpha} + \boldsymbol{n}) \left(\prod_{j\neq i}^{|\Omega|-1} \theta_j^{\nu\alpha_j + n_j - 1} \theta_{|\Omega|}^{\nu(1-\alpha_1-\ldots-\alpha_{i-1}+\alpha_{i+1}-\ldots-\alpha_{|\Omega|-1})+n_{|\Omega|}-1} \right. \right. \\
&\qquad \left. \left. \frac{\partial}{\partial \alpha_i} \theta_i^{\nu\alpha_i + n_i - 1} \theta_{|\Omega|}^{-\nu\alpha_i} \right) - \prod_{j=1}^{|\Omega|} \theta_j^{\nu\alpha_j + n_j - 1} \left(\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha} + \boldsymbol{n}) \right) \right) \\
&= \frac{1}{B(\nu\boldsymbol{\alpha}' + \boldsymbol{n})^2} \left(B(\nu\boldsymbol{\alpha}' + \boldsymbol{n}) \nu (\log\theta_i - \log\theta_{|\Omega|}) \prod_{j=1}^{|\Omega|} \theta_j^{\nu\alpha_j + n_j - 1} \right. \\
&\qquad \left. - \prod_{j=1}^{|\Omega|} \theta_j^{\nu\alpha_j + n_j - 1} \left(\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha}' + \boldsymbol{n}) \right) \right) \\
&= \nu(\log\theta_i - \log\theta_{|\Omega|}) p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}) - \frac{\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha}' + \boldsymbol{n})}{B(\nu\boldsymbol{\alpha}' + \boldsymbol{n})} p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}).
\end{aligned}
$$

This gradient is defined for each $\boldsymbol{\theta} \in \triangle^{\Omega}$ where $\boldsymbol{\theta} > \mathbf{0}$ when $\boldsymbol{n} > \mathbf{0}$. Then, for every such

fixed $\boldsymbol{\theta}$, we can apply the mean value theorem to find $\boldsymbol{\alpha}''$ such that,

$$
\begin{aligned}
&f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0) \\
&= \int_{\triangle^\Omega} |g| I(|g| > c)(\boldsymbol{\theta})(p_{\boldsymbol{\alpha}'}(\boldsymbol{\theta}) - p_{\boldsymbol{\alpha}'_0}(\boldsymbol{\theta}))d\boldsymbol{\theta} \\
&= \left(\int_{\triangle^\Omega} |g| I(|g| > c)(\boldsymbol{\theta}) \frac{\partial}{\partial \alpha_i} p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right)_i^T (\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0) \\
&= \left(\int_{\triangle^\Omega} |g| I(|g| > c)(\boldsymbol{\theta}) \left(\nu(\log \theta_i - \log \theta_{|\Omega|}) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) - \frac{\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu\boldsymbol{\alpha}'' + \boldsymbol{n})} p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) \right) d\boldsymbol{\theta} \right)_i^T (\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0).
\end{aligned}
$$

By Lemma B.3.6, the upper bound of the left integral is finite over the relevant variables:

$$
\begin{aligned}
K = \max_{i=1,\ldots,|\Omega|} \sup_{\boldsymbol{\alpha}'' \in \overline{\blacktriangle^{|\Omega|-1}}} \nu \Bigg(& \int_{\triangle^\Omega} (|g(\boldsymbol{\theta}) \log \theta_i| + |g \log \theta_{|\Omega|}|) I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
& - \frac{\frac{\partial}{\partial \alpha_i} B(\nu\boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu\boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \Bigg) < \infty.
\end{aligned}
$$

Then, in all,

$$|f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0)| \leq |K| \, ||\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0||_1,$$

where $|\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0| = (|\alpha_{1i} - \alpha_{2i} : \ i = 1, \ldots, \Omega)$. So, if the positive quantity ϵ is defined as,

$$|K| \, ||\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0||_1 = \epsilon,$$

then, the choice of $\delta_{\epsilon,\boldsymbol{\alpha}'_0}$ such that,

$$||\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0||_1 \leq \delta_{\epsilon,\boldsymbol{\alpha}'_0},$$

implies,

$$|f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0)| \leq \epsilon.$$

This is such that,

$$\forall \epsilon > 0, \ \exists \ \delta_{\epsilon,\boldsymbol{\alpha}'_0} : \ (||\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0||_1 \leq \delta_{\epsilon,\boldsymbol{\alpha}'_0} \ \Rightarrow \ |f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0)| \leq \epsilon.$$

Finally, because continuity is preserved over all p-norms,

$$\forall \epsilon > 0, \ \exists \ \delta_{\epsilon,\boldsymbol{\alpha}'_0} : \ (||\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_0||_p \leq \delta_{\epsilon,\boldsymbol{\alpha}'_0} \ \Rightarrow \ |f_c(\boldsymbol{\alpha}') - f_c(\boldsymbol{\alpha}'_0)| \leq \epsilon.$$

□

Lemma B.3.6: When $\boldsymbol{n} > \mathbf{0}$, the upper bound,

$$K = \max_{i=1,\ldots,|\Omega|} \sup_{\boldsymbol{\alpha}'' \in \blacktriangle^{|\Omega|-1}} \nu \left(\int_{\triangle^\Omega} (|g(\boldsymbol{\theta}) \log \theta_i| + |g \log \theta_{|\Omega|}|) I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right.$$
$$\left. - \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right).$$

is finite.

Proof: For every $i = 1, \ldots, |\Omega|$,

$$\int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) \left(\nu (\log \theta_i - \log \theta_{|\Omega|}) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) - \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) \right) d\boldsymbol{\theta}$$
$$= \nu \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| (\log \theta_i - \log \theta_{|\Omega|}) I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$- \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$\leq \nu \int_{\triangle^\Omega} |g(\boldsymbol{\theta}) (\log \theta_i - \log \theta_{|\Omega|})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$- \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$\leq \nu \int_{\triangle^\Omega} (|g(\boldsymbol{\theta}) \log \theta_i| + |g \log \theta_{|\Omega|}|) I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$- \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^\Omega} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Now, for any cell probability θ_k, by Hölder's inequality,

$$\int |g \log \theta_k| p_{\boldsymbol{\alpha}''} d\boldsymbol{\theta} \leq \left(\int g^2 p_{\boldsymbol{\alpha}''} d\boldsymbol{\theta} \right)^{1/2} \left(\int (\log \theta_k)^2 p_{\boldsymbol{\alpha}''} d\boldsymbol{\theta} \right)^{1/2},$$

and the right quantities are finite because the second moments finitely exist for both g and $\log \theta_k$ under the Dirichlet distribution of $p_{\boldsymbol{\alpha}''}$.

Also, the second term involving the derivative of the beta function is finite for every $\boldsymbol{\alpha}'' \in \overline{\blacktriangle^{|\Omega|-1}}$, since, by Lemma B.3.2, the log odds g is integrable and the digamma and trigamma function in the Beta function and its derivative are finite when $\boldsymbol{n} > \mathbf{0}$ over $\boldsymbol{\alpha}'' \in \overline{\triangle^{\Omega}}$.

Then, any sequence of evaluations of this expression over $\boldsymbol{\alpha}''$ in $\overline{\blacktriangle^{|\Omega|-1}}$ must also have a finite limit, meaning that the supremum over the closed set $\overline{\blacktriangle^{|\Omega|-1}}$ is finite. Let us denote this finite upper bound by,

$$K = \max_{i=1,\ldots,|\Omega|} \sup_{\boldsymbol{\alpha}'' \in \overline{\blacktriangle^{|\Omega|-1}}} \nu \left(\int_{\triangle^{\Omega}} (|g(\boldsymbol{\theta}) \log \theta_i| + |g \log \theta_{|\Omega|}|) I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right.$$
$$\left. - \frac{\frac{\partial}{\partial \alpha_i} B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})}{B(\nu \boldsymbol{\alpha}'' + \boldsymbol{n})} \int_{\triangle^{\Omega}} |g(\boldsymbol{\theta})| I(|g(\boldsymbol{\theta})| > c) p_{\boldsymbol{\alpha}''}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right).$$

□

B.3.4 Auxiliary results

In this appendix section, we make use of the following theorems from other authors.

Theorem B.3.4: (Yeh [87], Theorem 9.17, p.186) Let $(X, \mathcal{F}, \mu)$ be an arbitrary measure space. Let $(f_n : c \in \mathbb{N})$ be a monotone sequence of extended real-valued $\mathcal{F}$-measurable functions on a set $D \in \mathcal{F}$ and let $f = \lim_{c \to \infty} f_n$.

If $(f_c : c \in \mathbb{N})$ is a decreasing sequence, and there exists a μ-integrable extended real-valued $\mathcal{F}$-measurable function g such that $f_c \leq g$ on D for every $c \in \mathbb{N}$, then,

$$\lim_{c \to \infty} \int_D f_c d\mu = \int_D f d\mu.$$

□

Let us relax the sure convergence condition to almost sure.

Corollary B.3.1: Let $(X, \mathcal{F}, P)$ be an arbitrary probability space. Let $(f_n : c \in \mathbb{N})$ be a monotone sequence of extended real-valued $\mathcal{F}$-measurable functions on a set $D \in \mathcal{F}$ and

let $P(f = \lim_{c\to\infty} f_c) = 1$.

If $(f_c : c \in \mathbb{N})$ is a decreasing sequence, and there exists a P-integrable extended real-valued $\mathcal{F}$-measurable function g such that $f_c \leq g$ on D for every $c \in \mathbb{N}$, then,

$$\lim_{c\to\infty} \int_D f_c d\mu = \int_D f dP.$$

Proof: Let us denote by E the set of sample points over which f_c converges to f almost surely. Write

$$h_c = f_c I_E, \quad h = f I_E.$$

We have that, for all $c \in \mathbb{N}$,

$$\int_D h_c dP = \int_D f_c dP, \quad \text{and} \quad \int_D h dP = \int_D f dP. \tag{B.7}$$

Also, by Theorem B.3.4, because $h_c \leq g$ for the g that bounds f_c, we also know that,

$$\lim_{c\to\infty} \int_E h_c dP = \int_E h dP.$$

Because $E \subseteq D$ in order for the almost sure convergence to be defined, and $P(E) = 1$,

$$\int_D h_c dP = \int_E h_c dP, \quad \int_D h dP = \int_E h dP,$$

such that,

$$\lim_{c\to\infty} \int_D h_c dP = \int_D h dP. \tag{B.8}$$

Substituting (B.7) into (B.8) yields the required result.

□

Theorem B.3.5: (Rudin [67], Theorem 7.13, p. 150) Suppose K is compact and

(a) $\{f_c\}$ is a sequence of continuous functions on K,

(b) $\{f_c\}$ converges pointwise to a continuous function f on K,

(c) $f_c(x) \geq f_{n+1}(x)$ for all $x \in K$ and $n = 1, 2, 3, \ldots$.

Then $f_c \to f$ uniformly on K.

□

B.4 Results on IDM log-odds imprecision

Proposition B.4.1: for n and Δn such that $n_C \geq 1$ and $\Delta n_C \geq 0$,

$$\overline{P}_{\text{IDM}}[\log \boldsymbol{\theta}_C | \nu\alpha + n] \geq \overline{P}_{\text{IDM}}[\log \boldsymbol{\theta}_C | \nu\alpha + \boldsymbol{n} + \Delta n].$$

Proof: recall that the imprecision of the log-probability is given by,

$$\overline{P}_{\text{IDM}}[\log \boldsymbol{\theta}_C | \nu\alpha + n] = \psi(\nu + n_C) - \psi(n_C).$$

The change in imprecision between having observations n and $n + \Delta n$ is,

$$\begin{aligned}
&\overline{P}_{\text{IDM}}[\log \boldsymbol{\theta}_C | \nu\alpha + \boldsymbol{n} + \Delta n] - \overline{P}_{\text{IDM}}[\log \boldsymbol{\theta}_C | \nu\alpha + n] \\
&\quad = \psi(\nu + n_C + \Delta n_C) - \psi(n_C + \Delta n_C) - \psi(\nu + n_C) + \psi(n_C).
\end{aligned}$$

We write this expression in terms of the derivative ψ' by using the mean value theorem for when $n_C \geq 1$. In particular, when $\nu > \Delta n$, there exists $u \in (\nu + n_C, \nu + n_C + \Delta n_C)$ and $v \in (n_C, n_C + \Delta n_C)$,

$$\begin{aligned}
&(\psi(\nu + n_C + \Delta n_C) - \psi(n_C + \nu)) - (\psi(\Delta n_C + n_C) + \psi(n_C)) \\
&= \Delta n_C(\psi'(u) - \psi'(v)) \\
&< 0,
\end{aligned}$$

where the last inequality is because ψ' is a strictly decreasing function and that $u > v$. This proves the desired result for $\nu > \Delta n$. The case $\nu \leq \Delta n$ results in the inequality,

$$\nu(\psi'(a) - \psi'(b)) < 0,$$

with $a \in (\Delta n_C + n_C, \nu + n_C + \Delta n_C)$ and $b \in (n_C, n_C + \nu)$, and this proves the desired result for $\nu \leq \Delta n$.

□

Lemma B.4.1: For two random variables X and Y, and a conjugate pair of coherent expectations $\underline{E}, \overline{E}$ (not necessarily IDM,)

$$\overline{P}[X + Y] \leq \overline{P}[X] + \overline{P}[Y].$$

$$\overline{P}[-X] = \overline{P}[X].$$

By proof of induction, for finite collections $X_1, \ldots, X_r$ and $Y_1, \ldots, Y_s$,

$$\overline{P}\left[\sum_{i=1}^{r} X_i - \sum_{j=1}^{s} Y_j\right] \leq \sum_{i=1}^{r} \overline{P}[X] + \sum_{j=1}^{s} \overline{P}[Y].$$

Proof: For the first statement, because $\sup\{p(x) + p(y) : \ p \in M\} \leq \sup\{p(x) : \ p \in M\} + \sup\{p(y) : \ p \in M\}$, and $\inf\{p(x) + p(y) : \ p \in M\} \geq \inf\{p(x) : \ p \in M\} + \inf\{p(y) : p \in M\}$, imprecision satisfies the triangular inequality

$$\begin{aligned}
\overline{P}[X+Y] &= \overline{E}[X+Y] - \underline{E}[X+Y] \\
&\leq \overline{E}[X] + \overline{E}[Y] - \underline{E}[X] - \underline{E}[Y] \\
&\leq \overline{P}[X] + \overline{P}[Y].
\end{aligned}$$

For the second statement, suppose $\underline{E}$ and $\overline{E}$ for envelopes for a set of expectations M,

$$\begin{aligned}
\overline{P}[-X] &= \overline{E}[-X] - \underline{E}[-X] \\
&= \sup\{p(-X) : p \in M\} - \inf\{p(-X) : \ p \in M\} \\
&= \sup\{-p(X) : p \in M\} - \inf\{-p(X) : \ p \in M\} \\
&= -\inf\{p(X) : p \in M\} + \sup\{p(X) : \ p \in M\} \\
&= \overline{P}[X].
\end{aligned}$$

□

We now prove that the decrease of imprecision of the log-probability inference is greater when increasing the relevant event's count from a small number: in particular, the jump from no observation to a single observation is the greatest. Let us define the imprecision as well as the change of imprecision of the inference of a log probability $\log \boldsymbol{\theta}_C$ (for some set $C \subset \{1, \ldots, m\}$,

$$\begin{aligned}
\overline{P}(n) &:= \psi(\nu + n) - \psi(n), \\
\Delta\overline{P}(n; \Delta n) &:= \overline{P}(n + \Delta n) - \overline{P}(n).
\end{aligned}$$

In particular, the last expression is the change of imprecision at n when an additional $\Delta \boldsymbol{n} > 0$ counts are observed in the set C.

Proposition B.4.2: Let $n_1, n_2, \Delta n, \nu > 0$ such that $n_2 > n_1$. Then,

$$0 \geq \Delta \underline{\overline{P}}(n_2; \Delta n) > \Delta \underline{\overline{P}}(n_1; \Delta n).$$

Proof: From Lemma B.4.1, imprecision is a decreasing function so that the change of imprecision is a negative function.

For the next inequality, because sums of digamma functions is continuous and differentiable on the positive real line, we apply the mean value theorem to $n \in \mathbb{R} \longmapsto \Delta \underline{\overline{P}}(n; \Delta n) \in \mathbb{R}$ by,

$$\Delta \underline{\overline{P}}(n_2; \Delta n) - \Delta \underline{\overline{P}}(n_1; \Delta n) = \Delta \underline{\overline{P}}'(u; \Delta n)(n_2 - n_1),$$

with $u \in (n_1, n_2)$ and ,

$$\Delta \underline{\overline{P}}'(u; \Delta n) = \psi'(\nu + u + \Delta n) - \psi'(u + \Delta n) - \psi'(\nu + u) + \psi'(u).$$

Our goal is to show that $\Delta \underline{\overline{P}}'(u; \Delta n)$ is non-negative. Consider the case when $\nu > \Delta n$. Then, because the trigamma function, ψ', is again continuous and differentiable on the positive real line, we can apply the mean value theorem again to obtain,

$$\Delta \underline{\overline{P}}'(u; \Delta n) = (\psi'(\nu + u + \Delta n) - \psi'(\nu + u)) - (\psi'(u + \Delta \boldsymbol{n}) - \psi'(u)) = \Delta \boldsymbol{n} \psi''(w) - \Delta \boldsymbol{n} \psi''(v),$$

with $w \in (\nu + u, \nu + u + \Delta n)$ and $v \in (u, u + \Delta n)$ such that $v < w$ (for any $u \in (n_1, n_2)$.) Because the tetragamma function, ψ'', is strictly increasing, $v < w$ implies $\psi''(v) < \psi''(w)$. Thus, overall, we have, for u, v, w satisfying their respective bounds,

$$\Delta \underline{\overline{P}}'(u; \Delta n) = \Delta \boldsymbol{n}(n_2 - n_1)(\psi''(w) - \boldsymbol{n} \psi''(v)) > 0.$$

On the other hand, if $\nu \leq \Delta n$, one could instead use the difference,

$$\Delta \underline{\overline{P}}'(u; \Delta n) = (\psi'(\nu + u + \Delta n) - \psi'(u + \Delta n)) - (\psi'(\nu + u) - \psi'(u)) = \nu \psi''(a) - \nu \psi''(b),$$

with $a \in (\Delta \boldsymbol{n} + u, \nu + u + \Delta n)$ and $b \in (u, \nu + u)$ instead and apply the same procedure to obtain the same result. This completes the proof.

□

B.5 Indeterminate forms and their limiting processes

In the following, we demonstrate that when indeterminacy can happen in an expectation of log-odds under a Dirichlet distribution, then the expectation can take on any value of the extended real line. We do so by constructing a limiting process for each value it can obtain.

Theorem B.5.1: Suppose that $\bigcup_{i=1}^{r} A_i \cup \bigcup_{j=1}^{s} B_j \subset \{1, \ldots, m\}$ (i.e. it does not cover the whole set of indices.) For any $m_i, m_j \in \mathbb{R}^+ \cup \{0\}$, the finite collection of paths $\{\alpha_i : \ i = 1, \ldots, r\}$ and $\{\alpha_j : \ j = 1, \ldots, s\}$ satisfying,

$$\left(\sum_{i=1}^{r} \frac{1}{\alpha_i(t)} \right) = \log m_i t, \ \left(\sum_{j=1}^{s} \frac{1}{\alpha_j(t)} \right) = \log m_j t,$$

and, as $t \to \infty$, $\alpha_i(t), \alpha_j(t) \to 0$, will yield,

$$\sum_{i=1}^{r} \psi(\nu \alpha_i(t)) - \sum_{j=1}^{s} \psi(\nu \alpha_j(t)) \to \frac{1}{\nu} \log \frac{m_i}{m_j}.$$

Proof: The Laurent expansion of the digamma function around 0 implies that,

$$\psi(x) \approx -\frac{1}{x},$$

when x is in a sufficiently small neighbourhood of 0. Then,

$$\sum_{i=1}^{r} \psi(\nu \alpha_{A_i}) - \sum_{j=1}^{s} \psi(\nu \alpha_{B_j}) \approx -\frac{1}{\nu} \sum_{i=1}^{r} \frac{1}{\alpha_{A_i}} + \frac{1}{\nu} \sum_{j=1}^{s} \frac{1}{\alpha_{B_j}}.$$

This expression can be exponentiated to obtain,

$$\frac{1}{\nu} \log \frac{\exp\left(\sum_{i=1}^{r} \frac{1}{\alpha_{A_i}} \right)}{\exp\left(\sum_{j=1}^{s} \frac{1}{\alpha_{B_j}} \right)}.$$

Then, substituting the construction,

$$\left(\sum_{i=1}^{r} \frac{1}{\alpha_i(t)} \right) = \log m_i t, \ \left(\sum_{j=1}^{s} \frac{1}{\alpha_j(t)} \right) = \log m_j t,$$

yields the desired result.

□

Notice that the possibility of $m_i = 0$ and $m_j > 0$ yields $-\infty$ as a limit and $m_i > 0$ and $m_j = 0$ yields $+\infty$.

Using this result as well as the consideration of aggregation of boundary Dirichlet distributions, one can characterise the degeneracy of the expected general log odds when the observations are sparse in some cells involved in the general log odds. Indeed, consider,

$$E\left[\log \frac{\boldsymbol{\theta}_{A_1} \dots \boldsymbol{\theta}_{A_r}}{\boldsymbol{\theta}_{B_1} \dots \boldsymbol{\theta}_{B_s}} \,\middle|\, \nu\alpha + \boldsymbol{n}\right] = \Psi(\mathcal{A}, \mathcal{B}; \nu, \alpha, n) + \sum_{i=1}^{r} \psi(\nu\alpha_{A_i \cap I}) - \sum_{j=1}^{s} \psi(\nu\alpha_{B_j \cap I}),$$

where $\mathcal{A} = \{A_1, \dots, A_r\}$, $\mathcal{B} = \{B_1, \dots, B_s\}$ and Ψ is the term to contain the sums over the complements $A_i - I$ and $B_j - I$ (which are finite since their n's are non-zero.) The analysis can then be boiled down to the expected value tending to $+\infty$ on the face,

$$\left\{\alpha : \bigvee_{i=1}^{r} \alpha_{A_i \cap I} = 0 \wedge \bigwedge_{j=1}^{s} \alpha_{B_j \cap I} > 0\right\},$$

to $-\infty$ on the face,

$$\left\{\alpha : \bigwedge_{i=1}^{r} \alpha_{A_i \cap I} > 0 \wedge \bigvee_{j=1}^{s} \alpha_{B_j \cap I} = 0\right\},$$

and indeterminate forms that take on any finite real values on,

$$\left\{\alpha : \bigvee_{i=1}^{r} \alpha_{A_i \cap I} = 0 \wedge \bigvee_{j=1}^{s} \alpha_{B_j \cap I} = 0\right\}.$$

B.6 Properties of Dirichlet-multinomial conjugate pair

We briefly review the construction of a posterior distribution when the likelihood is induced by an observation model of an exponential family. We follow Bickis [14], whose construction generalises that of the usual interpretation put forth by Diaconis and Ylvisaker [36].

Let $x \in X$ be i.i.d. data observed from an exponential family observation model spanned by finite number of sufficient statistics $v = (v_1(x), \ldots, v_k(x))$ with natural parameters ξ. Then, the log-likelihood can be considered to be the logarithm of the Radon-Nikodym derivative of the posterior with respect to a prior measure Π_0 when Bayes' rule is applied:

$$\log \frac{d\Pi_x}{d\Pi_0}(\xi) = \sum_{i=1}^{m} \xi_i v_i(x) - 1\phi(\xi) =: \xi^T v(x),$$

$$\phi(\xi) = \log \int e^{\xi^T v(x)} dx,$$

(where $v(x) = (1, v_1, \ldots, v_k)$ and dx denotes an ambient measure on X.) The natural parameter space is the set $\Xi = \{\xi : \ \phi(\xi) \leq \infty\}$. Given this likelihood, one can define the exponential family on Ξ with sufficient statistics $v^*(\xi) = (-\phi(\xi), \xi_1, \ldots, \xi_k)$ and natural parameters $\eta = (\eta_0, \eta_1, \ldots, \eta_k)$, such that,

$$\log \frac{d\Pi_\eta}{d\xi}(\xi) = \eta^T v^* - \psi(\eta),$$

$$\psi(\eta) = \log \int e^{\eta^T v^*(\xi)} d\xi,$$

(where $d\xi$ denotes an ambient measure on Ξ.) The natural parameter space is $H = \{\eta : \psi(\eta) < \infty\}$. When the prior belongs to this exponential family with coordinates η_0, the posterior therefore has density with respect to $d\xi$ proportional to,

$$\exp(\xi^T v(x) + {\eta^{(0)}}^T v^*(\xi)) = \exp((\eta^{(0)} + v(x))^T v^*(\xi)).$$

Thus, the posterior is again in the family spanned by v^* over Ξ. In particular, we arrive at the following result for posterior updating in this conjugate system.

Because the Dirichlet distribution is conjugate to the finite dimensional multinomial likelihood, the prior and posterior imprecise probabilities can be cast into the framework above.

In particular,

Theorem B.6.1: the posterior set of distributions of the IDM with $m < \infty$ categories is,

$$M|x = \left\{\Pi_{\eta^{(0)}+v(x)} : \ \eta^{(0)} \in \overline{\triangle^m}\right\}.$$

□

A Dirichlet density on the multinomial probabilities $\boldsymbol{\theta} \in \overline{\triangle^m}$ with natural parameters $\nu\alpha$, with $\alpha \in \overline{\triangle^m}$, is proportional to the kernel,

$$\prod_{i=1}^{m-1} \boldsymbol{\theta}_i^{\nu\alpha_i - 1}(1 - \theta_1 - \ldots - \boldsymbol{\theta}_{m-1})^{\nu(1-\alpha_1-\ldots-\alpha_{m-1})-1},$$

which is equivalent to,

$$\exp\left(\sum_{i=1}^{m-1}(\nu\alpha_i - 1)\xi_i - (\nu - m)\log\left(1 + \sum_{i=1}^{m-1} e^{\xi_i}\right)\right),$$

under the following parametrisation,

$$\xi_i(\boldsymbol{\theta}) = \log\frac{\boldsymbol{\theta}_i}{1 - \theta_1 - \ldots - \boldsymbol{\theta}_{m-1}}.$$

Therefore, we have the following lemma.

Lemma B.6.1: a Dirichlet distribution with parameters $\nu\alpha$ is identified with natural coordinates/parameters,

$$\eta = (\nu, \nu\alpha_1, \ldots, \nu\alpha_{m-1}) + (-m, -1, \ldots, -1),$$

relative to the span,

$$v^*(\xi) = \left(-\log\left(1 + \sum_{i=1}^{m-1} e^{\xi}\right), \xi_1, \ldots, \xi_{m-1}\right).$$

□

The multinomial likelihood can be similarly cast: with probabilities $\boldsymbol{\theta} \in \overline{\triangle^m}$, it is proportional to the kernel,

$$\prod_{i=1}^{m-1} \boldsymbol{\theta}_i^{n_i} (1 - \theta_1 - \ldots - \boldsymbol{\theta}_{m-1})^{n_k},$$

where $(n_1, \ldots, n_k)$ are the observation counts which sum to $|n| = 1^T n$. Its natural exponential form is,

$$\sum_{i=1}^{m-1} n_i \xi_i - |n| \log \left(1 + \sum_{i=1}^{m-1} e^{\xi_i} \right),$$

under the following parametrisation

$$\xi_i(\boldsymbol{\theta}) = \log \frac{\boldsymbol{\theta}_i}{1 - \theta_1 - \ldots - \boldsymbol{\theta}_{m-1}}.$$

Therefore, we have the following lemma.

Lemma B.6.2: a (finite dimensional) multinomial distribution with probabilities $\boldsymbol{\theta} \in \triangle^m$ is identified with natural coordinates/parameters,

$$\xi = \left(-\log \left(1 + \sum_{i=1}^{m-1} e^{\xi_i} \right), \xi_1, \ldots, \xi_{m-1} \right),$$

relative to the span,

$$v(n) = (|n|, n_1, \ldots, n_{m-1}).$$

□

Appendix C

Appendix to Chapter 4

C.1 Optimisation algorithm used in Section 4.7

All optimisations for lower and upper IDM expectations in Section 4.7 are done using a genetic algorithm built from the `deap` framework (De Raiville et al. [33]) under `Python3.5`.

Algorithm 1: Genetic algorithm for IDM optimisation

```
# Loop over generations
population = initialise_population();
while generation < max_iteration do
    offsprings = ∅
    # Select survivours by tournament
    fitnesses = evaluate_fitness(population)
    survivours = tournament(population, fitnesses)
    # Cross over survivours and create offsprings
    for surv1 ≠ surv2 in survivours_without_replacement do
        if Uniform(0,1) < crossover_prob then
            offspring1, offspring2 = two_point_crossover(surv1, surv2)
        else
            offspring1, offspring2 = surv1, surv2
        end
        offsprings.append(offspring1)
        offsprings.append(offspring2)
    end
    # Mutate offsprings
    for offspring in offsprings do
        if Uniform(0,1) < mutation_prof then
            offspring = mutate(offspring)
        end
    end
    # Assign offsprings to population of next generation.
    population = offprings
end
```

A population of 1000 individuals are intialised using a flat Dirichlet distribution. and the cross-over probability and the probability for bit mutation of each individual are 0.5 and 0.2, respectively. Where the optimisation is constrained over $\overline{\triangle^m}$, the objective function is first penalised by the squared deviation of the sum of the point from unity,

$$\boldsymbol{\alpha} \longmapsto 1000(\mathbf{1} \cdot \boldsymbol{\alpha} - 1)^2.$$

Appendix D

Appendix to Chapter 5

D.1 Results and proofs

Lemma D.1.1: Let ψ be the digamma function over the domain of positive arguments. Then,

$$\psi^{-1\prime}(x) = \frac{1}{\psi'(\psi^{-1}(x))}.$$

Proof: The inverse derivative of ψ is justified as it is bijective and continuously differentiable: taking the derivative on both sides of,

$$x = \psi(\psi^{-1}(x)),$$

yields the result.

□

Lemma D.1.2: Let ψ be the digamma function over the domain of positive arguments. Then,

$$\psi^{-1\prime\prime}(x) = -\frac{\psi''(\psi^{-1}(x))}{(\psi'(\psi^{-1}(x)))^3}.$$

Proof: The inverse derivative of ψ and ψ' are justified as they are continuously differentiable. Then,

$$\begin{aligned}
\psi^{-1\prime\prime}(x) &= \frac{d}{dx}\psi^{-1\prime}(x) \\
&= \frac{d}{dx}\frac{1}{\psi'(\psi^{-1}(x))} && \text{(Lemma D.1.1)} \\
&= -(\psi'(\psi^{-1}(x)))^{-2}\cdot\psi''(\psi^{-1}(x))\cdot\frac{d}{dx}\psi^{-1}(x) \\
&= -(\psi'(\psi^{-1}(x)))^{-2}\cdot\psi''(\psi^{-1}(x))\cdot\frac{1}{\psi'(\psi^{-1}(x))} && \text{(Lemma D.1.1)} \\
&= -\frac{\psi''(\psi^{-1}(x))}{\psi'(\psi^{-1}(x))^3}.
\end{aligned}$$

□

Lemma D.1.3: Let ψ be the digamma function over the domain of positive arguments. Then, for the function,

$$I:\ \boldsymbol{\mu} \longmapsto \psi(\psi^{-1}(\mu_1)+\psi^{-1}(\mu_2)) - (\psi^{-1}(\mu_2)+\psi^{-1}(\mu_3)),$$

subject to the constraint,

$$\psi^{-1}(\mu_1)+\psi^{-1}(\mu_2)+\psi^{-1}(\mu_3) = \nu + n,$$

with $\nu + n > 0$, we have,

$$\nabla I > 0,$$

and I has no turning points over $\boldsymbol{\mu}$.

Proof: the gradient of I is given by,

$$\nabla I = \psi'(\psi^{-1}(\mu_1)+\psi^{-1}(\mu_2))\begin{bmatrix}\psi^{-1}(\mu_1)\\ \psi^{-1}(\mu_2)\\ 0\end{bmatrix} - \psi'(\psi^{-1}(\mu_2)+\psi^{-1}(\mu_3))\begin{bmatrix}0\\ \psi^{-1}(\mu_2)\\ \psi^{-1}(\mu_3)\end{bmatrix}.$$

Also, taking the gradient of the constraint equation yields,

$$[\partial_{\mu_1}\psi^{-1\prime}(\mu_1), \partial_{\mu_1}\psi^{-1\prime}(\mu_2), \partial_{\mu_1}\psi^{-1\prime}(\mu_3)] = \mathbf{0},$$

such that, substituting into the gradient of I,

$$\nabla I = \psi'(\psi^{-1}(\mu_1) + \psi^{-1}(\mu_2)) \begin{bmatrix} \psi^{-1}(\mu_1) \\ \psi^{-1}(\mu_2) \\ 0 \end{bmatrix} + \psi'(\psi^{-1}(\mu_2) + \psi^{-1}(\mu_3)) \begin{bmatrix} \psi^{-1}(\mu_1) \\ 0 \\ 0 \end{bmatrix}.$$

Now, ψ' is positive over positive arguments and ψ^{-1} maps any real number back to the positive line, so $\psi'(\psi'(\mu_j))$ is positive for $j = 1, 2, 3$. Furthermore, by Lemma D.1.1,

$$\psi^{-1\prime}(x) = \frac{1}{\psi'(\psi^{-1}(x))},$$

for all x, such that it is again positive by similar arguments. So, all elements of ∇I are positive. Because ∇I is continuous over $\boldsymbol{\mu}$ and strictly positive, it never changes signs and so there is no turning point.

□

Lemma D.1.4: For $\boldsymbol{t} = (t_1, t_2)$, $\boldsymbol{v}_0 = \mathbf{1}$ and a choice of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ with $\boldsymbol{v}_i \in \mathbb{R}^3$, the function a implicitly defined to satisfy,

$$\boldsymbol{\mu}(\boldsymbol{t}) = a(\boldsymbol{t})\boldsymbol{v}_0 + t_1\boldsymbol{v}_1 + t_2\boldsymbol{v}_2,$$

and,

$$\psi^{-1}(\mu_1(\boldsymbol{t})) + \psi^{-1}(\mu_2(\boldsymbol{t})) + \psi^{-1}(\mu_3(\boldsymbol{t})) = \nu + n,$$

is concave:

$$[\partial_{t_k,t_l} a(\boldsymbol{t})]_{k,l},$$

is negative definite.

Proof: For brevity, let us write,

$$x_{kj}(\boldsymbol{t}) := \partial_{t_k} a(\boldsymbol{t}) + \delta_{1k} v_{1j} + \delta_{2k} v_{2j}.$$

and similarly define $x_{lj}(\boldsymbol{t})$. Then, taking the partial derivative ∂_{t_l} of the second condition and substituting the first condition for $\boldsymbol{\mu}$ yield,

$$\sum_{j=1}^{3} \psi^{-1\prime}(\mu_j(\boldsymbol{t}))x_{lj}(t) = 0.$$

Applying a second partial derivative ∂_{t_k} yields,

$$\sum_{j=1}^{3} \psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))x_{lj}(\boldsymbol{t})x_{kj}(\boldsymbol{t}) + \psi^{-1\prime}(\mu_j(\boldsymbol{t}))\partial_{t_k,t_l}a(\boldsymbol{t}) = 0.$$

This solves for,

$$\partial_{t_k,t_l}a(\boldsymbol{t}) = -\frac{\sum_{j=1}^{3} \psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))x_{lj}(\boldsymbol{t})x_{kj}(\boldsymbol{t})}{\sum_{j=1}^{3} \psi^{-1\prime}(\mu_j(\boldsymbol{t}))}.$$

Notice that, by Lemma D.1.2,

$$\psi^{-1\prime\prime}(\mu_j) = -\frac{\psi''(\psi^{-1}(\mu_j))}{(\psi'(\psi^{-1}(\mu_j)))^3},$$

such that, because ψ'' is a negative function over positive arguments, ψ' is a positive function over positive arguments, and $\psi^{-1})(\mu_j)$ is always positive as ψ is bijective, continuous and restricted to the domain of positive real numbers,

$$\psi^{-1\prime\prime}(\mu_j) \geq 0.$$

The negative definiteness can now be shown as follows. Since $\psi^{-1\prime\prime}(\mu_j) \geq 0$, its square root is again a non-negative number such that we can rewrite,

$$\partial_{t_k,t_l}a(\boldsymbol{t}) = -\frac{\sum_{j=1}^{3}(\sqrt{\psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))}x_{lj}(\boldsymbol{t}))(\sqrt{\psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))}x_{kj}(\boldsymbol{t}))}{\sum_{j=1}^{3} \psi^{-1\prime}(\mu_j(\boldsymbol{t}))}.$$

Thus, the numerator is an element of a square product of a square matrix, which is positive definite. From Lemma D.1.1 and the fact that ψ' is a positive function over its domain, the denominator is a positive scalar constant, so the matrix,

$$[\partial_{t_k,t_l}a(\boldsymbol{t})]_{k,l} = -\left[\frac{\sum_{j=1}^{3}(\sqrt{\psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))}x_{lj}(\boldsymbol{t}))(\sqrt{\psi^{-1\prime\prime}(\mu_j(\boldsymbol{t}))}x_{kj}(\boldsymbol{t}))}{\sum_{j=1}^{3} \psi^{-1\prime}(\mu_j(\boldsymbol{t}))}\right]_{k.l},$$

is negative definite.

□

D.2 Some properties of the digamma function

The *digamma function* is defined as,

$$\psi(x) := \frac{d}{dx} \log \Gamma(x).$$

In this document, the digamma function is usually evaluated in the form $\psi\left(\nu \sum_{i \in A} \alpha_I\right)$ where $0 \leq \sum_{i \in A} \alpha_i \leq 1$. Indeed, it can be shown that (e.g. see [56]) over $x > 0$,

$$\psi(x) = -\gamma_0 + \sum_{n=0}^{\infty} \left(\frac{1}{1+n} - \frac{1}{x+n} \right),$$

where $\gamma_0 > 0$ is the Euler-Mascheroni constant.

We note the following properties of ψ.

Corollary D.2.1: From this expansion, ψ is a monotonically increasing function in x over $x > 0$.

□

Corollary D.2.2: The unique root x_0 of $\psi(x_0) = 0$ is greater than one. Indeed, $\psi(1) = -\gamma_0 < 0$ and so by increasing monotonicity, the root is unique and ψ must achieve zero above $x = 1$.

□

Corollary D.2.3: When $x \to \infty$, $\psi(x) \to \infty$. This happens as $\frac{1}{1+n} - \frac{1}{x+n} \approx \frac{1}{1+n} = O(1/n)$ and the infinite sum of the latter diverges.

□

A loose approximation of the root of $\psi(x) = 0$: Using the `scipy.special.digamma` (`scipy==1.1.0`) function in `python3.5`, we observe the loose bounds $\psi(1.25) < 0 < \psi(1.5)$ containing the root.

Appendix E

Appendix to Chapter 7

E.1 Results and proofs

Lemma E.1.1: (6.5.3 of Walley [81]) Let $\mathcal{B}$ be a partition of Ω. Let $\underline{E}$ be a lower expectation. For each $B \in \mathcal{B}$, let $\underline{E}(\cdot|B)$ be a lower expectation conditional on B. Write,

$$\underline{E}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} I_B \underline{E}(X|B).$$

Finally, let $\underline{E}$ and $\underline{E}(\cdot|B)$ for each $B \in \mathcal{B}$ be defined on $\mathcal{L}(\Omega)$. Then, $\underline{E}$ and $\underline{E}(\cdot|\mathcal{B})$ are jointly coherent iff,

- for all $X \in \mathcal{L}(\Omega)$, $\underline{E}(X - \underline{E}(X|\mathcal{B})) \geq 0$, and
- for all $X \in \mathcal{L}(\Omega)$ and $B \in \mathcal{B}$, $\underline{E}(I_B(X - \underline{E}(X|B))) = 0$.

□

Lemma E.1.2: (2.6.1 of Walley [81]) Let $\underline{E}$ be a coherent lower expectation. Then, for $X \geq Y$, $\underline{E}(X) \geq \underline{E}(Y)$.

□

For every $X \in \mathcal{L}(\Omega)$, write,

$$\underline{R}(X|\mathcal{B}) := \sum_{B \in \mathcal{B}} I_B \underline{R}(X|B).$$

Theorem E.1.1: Let $\mathcal{B}$ be a partition of Ω, and M be a closed set of distributions. Suppose that N is such that $\underline{E}^{\uparrow}_N(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$. Suppose also that the regular extension $\underline{R}_M(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$. Then, $\underline{E}_{M;N}(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}_M$.

Proof: First, notice that,

$$\underline{E}_{M;N}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} I_B \min\left\{\underline{R}_M(X|B), \underline{E}^{\uparrow}_N(X|B)\right\} \leq \sum_{B \in \mathcal{B}} I_B \underline{R}_M(X|B) = \underline{R}_M(X|\mathcal{B}),$$

and,

$$\underline{E}_{M;N}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} I_B \min\left\{\underline{R}_M(X|B), \underline{E}^{\uparrow}_N(X|B)\right\} \leq \sum_{B \in \mathcal{B}} I_B \underline{E}^{\uparrow}_N(X|B) = \underline{E}^{\uparrow}_N(X|\mathcal{B}).$$

This leads to,

$$X - \underline{E}_{M;N}(X|\mathcal{B}) \geq X - \min\left\{\underline{R}_M(X|\mathcal{B}), \underline{E}^{\uparrow}_N(X|\mathcal{B})\right\},$$

and, in turn, because

$$X - \min\left\{\underline{R}_M(X|\mathcal{B}), \underline{E}^{\uparrow}_N(X|\mathcal{B})\right\} \geq X - \underline{E}^{\uparrow}_N(X|\mathcal{B}),\ X - \underline{R}_M(X|\mathcal{B}),$$

by Lemma E.1.2,

$$\underline{E}_M\left(X - \underline{E}_{M;N}(X|\mathcal{B})\right) \geq \underline{E}_M(X - \underline{E}^{\uparrow}_N(X|\mathcal{B})),\ \underline{E}_M(X - \underline{R}_M(X|\mathcal{B})).$$

By Lemma E.1.1 and that both $\underline{R}_M(\cdot|\mathcal{B})$ and $\underline{E}^{\uparrow}_N(\cdot|\mathcal{B})$ are both jointly coherent with $\underline{E}_M$, we have that, $\underline{E}(X - \underline{R}_M(X|\mathcal{B})) \geq 0$ and $\underline{E}(X - \underline{E}^{\uparrow}_N(X|\mathcal{B})\} \geq 0$. This implies that,

$$\underline{E}(X - \underline{E}_{M;N}(X|\mathcal{B})) \geq \min\{\underline{E}(X - \underline{R}_M(X|\mathcal{B})), \underline{E}(X - \underline{E}^{\uparrow}_N(X|\mathcal{B})\} \geq 0. \tag{E.1}$$

Second, consider a fixed $B \in \mathcal{B}$ and a fixed $X \in \mathcal{L}(\Omega)$. Then,

$$\underline{E}_{M;N}(X|B) = \min\left\{\underline{R}_M(X|B), \underline{E}^{\uparrow}_N(X|B)\right\}$$

is a non-random constant whose value is either one of the real numbers in $\{\underline{R}_M(X|B), \underline{E}_{M;N}(X|B)\}$. Then, by Lemma E.1.1 and that $\underline{R}_M(\cdot|\mathcal{B})$ and $\underline{E}^{\uparrow}_N(\cdot|\mathcal{B})$ are both jointly coherent with $\underline{E}$, that,

$$\underline{E}(I_B(X - \underline{R}_M(X|B))) = 0,$$
$$\underline{E}(I_B(X - \underline{E}^{\uparrow}_N(X|B))) = 0,$$

we have,

$$\underline{E}(I_B(X - \underline{E}_{M;N}(X|B))) = 0. \tag{E.2}$$

By Lemma E.1.1, (E.1) and (E.2) imply that $\underline{E}_{M;N}(\cdot|\mathcal{B})$ is jointly coherent with $\underline{E}$.

□

Glossary

Bayes' rule – The property of expectations that the right hand side of $E_P(X|B) = E_P(I_B X)/P(B)$ exists when $P(B) > 0$, and uniquely relates $E_P(\cdot|B)$ and E_P. (It does not refer to the relationship between the two conditional measures, $P(B|A)$ and $P(A|B)$.)

Coherence – Qualitatively, a self-consistency amongst assessments over a set of random variables such that they do not contradict one another. See Chapter 2.

Generalised Bayes' rule, GBR – the extension of Bayes' rule (c.f. *Bayes' rule*) to the imprecise case using expectations.

IDM – Abbreviation for 'imprecise Dirichlet model'.

Imprecise model – A model that is not a precise model. E.g. a lower expectation induced by a non-singleton set of distributions.

Imprecise probabilities/expectations – A special case of an imprecise model induced by forming the lower and upper bounds of probabilities/expectations over a set of distributions. The term refers to the pair of lower and upper expectations used to mathematically represent them.

Precise model – A model that assigns a single fair price to every random variable in its domain. E.g. a single probability distribution assigning a single expectation value to all random variables when defined.

Vacuity, vacuous – The property that an imprecise model is non-informative regarding certain statistics of interest. For example, a pair of lower and upper imprecise expectations of a random variable X is considered vacuous (for X) if the lower and

upper bounds are trivially the infimum and the supremum of the random variable, respectively.