Journal of Artificial Intelligence Research 52 (2015) 1-95 Submitted 07/14; published 01/15

Coherent Predictive Inference under Exchangeability with Imprecise Probabilities

Gert de Cooman GERT.DECOOMAN@UGENT.BE
Jasper De Bock JASPER.DEBOCK@UGENT.BE
Ghent University, SYSTeMS Research Group
Technologiepark–Zwijnaarde 914
9052 Zwijnaarde, Belgium

Márcio Alves Diniz MARCIO.ALVES.DINIZ@GMAIL.COM
Federal University of São Carlos, Department of Statistics
Rod. Washington Luis, km 235
São Carlos, Brazil

Abstract

Coherent reasoning under uncertainty can be represented in a very general manner by coherent sets of desirable gambles. In a context that does not allow for indecision, this leads to an approach that is mathematically equivalent to working with coherent conditional probabilities. If we do allow for indecision, this leads to a more general foundation for coherent (imprecise-)probabilistic inference. In this framework, and for a given finite category set, coherent predictive inference under exchangeability can be represented using Bernstein coherent cones of multivariate polynomials on the simplex generated by this category set. This is a powerful generalisation of de Finetti's Representation Theorem allowing for both imprecision and indecision.

We define an inference system as a map that associates a Bernstein coherent cone of polynomials with every finite category set. Many inference principles encountered in the literature can then be interpreted, and represented mathematically, as restrictions on such maps. We discuss, as particular examples, two important inference principles: representation insensitivity—a strengthened version of Walley's representation invariance—and specificity. We show that there is an infinity of inference systems that satisfy these two principles, amongst which we discuss in particular the skeptically cautious inference system, the inference systems corresponding to (a modified version of) Walley and Bernard's Imprecise Dirichlet Multinomial Models (IDMM), the skeptical IDMM inference systems, and the Haldane inference system. We also prove that the latter produces the same posterior inferences as would be obtained using Haldane's improper prior, implying that there is an infinity of proper priors that produce the same coherent posterior inferences as Haldane's improper one. Finally, we impose an additional inference principle that allows us to characterise uniquely the immediate predictions for the IDMM inference systems.

1. Introduction

This paper deals with predictive inference for categorical variables. We are therefore concerned with a (possibly infinite) sequence of variables X_n that assume values in some finite set of categories A. After having observed a number $\check{n}$ of them, and having found that, say $X_1 = x_1$, $X_2 = x_2$, $\dots$, $X_{\check{n}} = x_{\check{n}}$, we consider some subject's belief model for the next $\hat{n}$ variables $X_{\check{n}+1}, \dots X_{\check{n}+\hat{n}}$. In the probabilistic tradition—and we want to build on this tradition in the context of this paper—this belief can be modelled by a conditional predictive probability mass function $p^{\hat{n}}(\cdot|x_1,\dots,x_{\check{n}})$ on the set $A^{\hat{n}}$ of their possible values. These probability mass functions can be used for prediction or estimation, for statistical inferences, and in decision making involving the uncertain values of these variables. In this sense, predictive inference lies at the heart of statistics, and more generally, of learning under uncertainty. For this reason, it is also of crucial importance for dealing with uncertainty in Artificial Intelligence, where for instance, intelligent systems have to learn about multinomial probabilities, or Markov transition probabilities, rates of occurrence for phenomena, local probabilities in Bayesian or credal networks and so on. We refer to the synthesis by Geisser (1993) and the collection of essays by Zabell (2005) for good introductions to predictive inference and the underlying issues that the present paper will also be concerned with.

What connects these predictive probability mass functions for various values of $\check{n}$, $\hat{n}$ and $(x_1,\dots,x_{\check{n}})$ are the requirements of *time consistency* and *coherence*. The former requires that when $n_1 \leq n_2$, then $p^{n_1}(\cdot|x_1,\dots,x_{\check{n}})$ can be obtained from $p^{n_2}(\cdot|x_1,\dots,x_{\check{n}})$ through the usual marginalisation procedure; while the latter essentially demands that these conditional probability mass functions should be connected with time-consistent unconditional probability mass functions through Bayes's Rule.

A common assumption about the variables X_n is that they are *exchangeable*, meaning roughly that the subject believes that the order in which they are observed, or present themselves, has no influence on the decisions and inferences he will make regarding these variables. This assumption, and the analysis of its consequences, goes back to de Finetti (1937) (see also Cifarelli & Regazzini, 1996). His famous Representation Theorem states, in essence, that the time-consistent and coherent conditional and unconditional predictive probability mass functions associated with a countably infinite exchangeable sequence of variables in A are completely characterised by[1]—and completely characterise—a unique probability measure on the Borel sets of the simplex of all probability mass functions on A, called their *representation*.[2]

This leads us to the central problem of predictive inference: since there is an infinity of such probability measures on the simplex, which one does a subject choose in a particular context, and how can a given choice be motivated and justified? The subjectivists of de Finetti's persuasion might answer that this question needs no answer: a subject's personal predictive probabilities are entirely his, and time consistency and coherence are the only requirements he should heed. Earlier scholars, like Laplace and Bayes, whom we would now also call subjectivists, invoked the Principle of Indifference to justify using a specific class of predictive mass functions. Proponents of the logicist approach to predictive inference would try enunciating general inference principles in order to narrow down, and hopefully eliminate entirely, the possible choices for the representing probability measures on the simplex. The logicians W. E. Johnson (1924) and, in a much more systematic fashion, Rudolf Carnap (1952)

1. … unless the observed sequence has probability zero.
2. Actually, in order to clarify the connection with what we shall do later on, the essence of de Finetti's argument is that the representation is a coherent prevision on the set of all multinomial polynomials—or equivalently, of all continuous real functions—on this simplex (De Cooman, Quaeghebeur, & Miranda, 2009b). As a (*finitely* additive) coherent prevision, it can be extended uniquely only so far as to the set of all lower semicontinuous functions, but it does determine a unique (*countably* additive) probability measure on the Borel sets of that simplex, through the F. Riesz Representation Theorem (De Cooman & Miranda, 2008a; Troffaes & De Cooman, 2014).

tried to develop an axiom system for predictive inference based on such reasonable inference principles. Carnap's first group of axioms is related to what we have called coherence, but as we suggested, these by themselves are too weak to single out a particular predictive model. His second group consisted of invariance axioms, including exchangeability. He also included an axiom of instantial relevance, translating the intuitive principle that predictive inferences should actually learn from experience. His last axiom, predictive irrelevance, was also proposed earlier by Johnson and called the *sufficientness postulate* by Good (1965). Armed with these axioms, Carnap was able to derive a continuum of probabilistic inference rules, closely related to the Dirichlet multinomial model and to the Imprecise Dirichlet Multinomial Model (IDMM) proposed by Walley (1996) and Walley and Bernard (1999), which we discuss in Appendices C and D, respectively.

Our point of view holds the middle ground between the subjectivist and logicist positions: it should be possible for a subject to make assessments for certain predictive probabilities, and to combine these with certain inference principles he finds reasonable, or which suit his purpose for the problem at hand. Indeed, the inference systems we introduce and discuss in Section 6, and the notion of conservative coherent inference—or natural extension—we associate with them, provide an elegant framework and tools for making conservative coherent predictive inferences that combine (local) subjective probability assessments with (general) inference principles. And our work in Section 15 on characterising the immediate predictions for the IDMM constitutes an exercise in—or an example for—precisely that.

This idea of *conservative probabilistic inference* brings us to what we believe is the main contribution of this paper. It is a central idea in de Finetti's (1975) approach to probability—but also of course implicit in the Markov and Chebyshev inequalities—that when a subject makes probability assessments, we can consider them as bounds on so-called precise probability models. Calculating such most conservative but tightest bounds is indeed what de Finetti's (1975) Fundamental Theorem of Prevision (see also Lad, 1996) is about. The theory of imprecise probabilities, brought to a synthesis by Williams (1976) and Walley (1991, 2000), but going back to Boole (1952) and Keynes (1921), with crucial contributions by quite a number of statisticians and philosophers (Smith, 1961; Levi, 1980; Seidenfeld, Schervish, & Kadane, 1999), looks at conservative probabilistic inference precisely in this way: how can we calculate as efficiently as possible the consequences—in the sense of most conservative tightest bounds—of making certain probability assessments. These may be local assessments, such as inequalities imposed on the probabilities or previsions of certain events or variables, or structural assessments, such as independence, or exchangeability.

One advantage of imprecise probability models is that they allow for *imprecision*, or in other words, the use of *partial* probability assessments using bounding *inequalities* rather than equalities. Another, related, advantage is that they allow for *indecision* to be modelled explicitly: loosely stated, if the imposed bounds on probabilities allow for more than one probability model as a solution, it may very well be that of two actions, the first has the higher expected utility for one compatible probability model, and the smaller for another compatible probability model, meaning that neither action is robustly preferred over the other. So with this current stated model for his beliefs, a subject is then undecided between these actions. In Section 2, we give a concise overview of the relevant ideas, models and techniques in the field of imprecise probabilities. A much more extensive and detailed recent overview of this area of research was published by Augustin, Coolen, De Cooman, and Troffaes (2014).

The present paper, then, can be described as an application of ideas in imprecise probabilities to predictive inference. Its aim is to study—and develop a general framework for dealing with—conservative coherent predictive inference using imprecise probability models. Using such models will also allow us to represent a subject's indecision, which we believe is a natural state to be in when knowing, or having learned little, about the problem at hand. It seems important that theories of learning under uncertainty in general, and predictive inference in particular, at least allow us (i) to start out with conservative, very imprecise and indecisive models when little has been learned, and (ii) to become more precise and decisive as more observations come in. We shall see that the abstract notion of an inference system that we introduce further on, allows for—but does not necessarily force—such behaviour, and we shall give a number of examples of concrete inference systems that display it.

Our work here builds on, but manages to reach much further than, an earlier paper by one of the authors (De Cooman, Miranda, & Quaeghebeur, 2009a). One reason why it does so, is that this earlier work deals only with immediate prediction models, and as we shall see further on, *predictive inference using imprecise probabilities is not completely determined by immediate prediction*, contrary to what we can expect when using precise probabilities. But the main reason is that we are now in a position to use a very powerful mathematical language to represent imprecise-probabilistic inferences: Walley's (2000) coherent sets of desirable gambles. Earlier imprecise probability models (Boole, 1952, 1961; Koopman, 1940) centred on lower and upper probability bounds for events—or propositions. Later on (Walley, 1991, Section 2.7), it became apparent that this language of events and *lower* and *upper probabilities* is lacking in power of expression: a much more expressive theory uses random variables and their *lower previsions* or *expectations*. This successful theory of coherent lower previsions is by now quite well developed (Walley, 1991; Augustin et al., 2014; Troffaes & De Cooman, 2014). But it faces a number of problems, such as its mathematical as well as conceptual complexity, especially when dealing with conditioning and independence, and the fact that, as is the case with many other approaches to probability, and as we shall see further on in Section 2.5, it has issues with conditioning on sets of (lower) probability zero.

A very attractive solution to these problems was offered by Walley (2000), in the form of coherent sets of desirable gambles, inspired by earlier ideas (Smith, 1961; Williams, 1975b; Seidenfeld, Schervish, & Kadane, 1995). Here, the primitive notions are not probabilities of events, nor expectations of random variables. The focus is rather on whether a gamble, or a risky transaction, is desirable to a subject—strictly preferred to the zero transaction, or status quo. And a basic belief model is now not a probability measure or lower prevision, but a *set of desirable gambles*. Of course, stating that a gamble is desirable also leads to a particular lower prevision assessment: it provides a lower bound of zero on the prevision of the gamble. We explain why we prefer to use sets of desirable gambles as basic uncertainty models in Section 2.

In summary, then, our aim in this paper is to use sets of desirable gambles to extend the existing probabilistic theory of predictive inference. Let us explain in some detail how we intend to go about doing this. The basic building blocks are introduced in Sections 2–7. As already indicated above, we give an overview of relevant notions and results concerning our imprecise probability model of choice—coherent sets of desirable gambles—in Section 2.

In particular, we explain how to use them for conservative inference as well as conditioning; how to derive more commonly used models, such as lower previsions and lower probabilities, from them; and how they relate to precise probability models.

In Section 3, we explain how we can describe a subject's beliefs about a sequence of variables in terms of predictive sets of desirable gambles, and the derived notion of predictive lower previsions. These imprecise probability models generalise the above-mentioned predictive probability mass functions $p^{\hat{n}}(\cdot|x_1, \ldots, x_{\check{n}})$, and they constitute the basic tools we shall be working with. We also explain what are the proper formulations for the above-mentioned time consistency and coherence requirements in this more general context.

In Section 4, we discuss a number of inference principles that we believe could be reasonably imposed on predictive inferences, and we show how to represent them mathematically in terms of predictive sets of desirable gambles and lower previsions. Pooling invariance—or what Walley (1996) has called the Representation Invariance Principle (RIP)—and renaming invariance seem reasonable requirements for any type of predictive inference, and category permutation invariance seems a natural thing to require when starting from a state of complete ignorance. Taken together, they constitute what we call *representation insensitivity*. It means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation—the category set. Another inference principle we look at imposes the so-called *specificity* property: when predictive inference is specific, then for a certain type of question involving a restricted number of categories, a more general model can be replaced by a more specific model that deals only with the categories of interest, and will produce the same relevant inferences (Bernard, 1997).

The next important step is taken in Section 5, where we recall from the literature (De Cooman et al., 2009b; De Cooman & Quaeghebeur, 2012) how to deal with exchangeability when our predictive inference models are imprecise. We recall that de Finetti's Representation Theorem can be significantly generalised. In this case, the time-consistent and coherent predictive sets of desirable gambles are completely characterised by a set of (multivariate) polynomials on the simplex of all probability mass functions on the category set.[3] This set of polynomials must satisfy a number of properties, which taken together define the notion of *Bernstein coherence*. Without becoming too technical at this point, the conclusion of this section is that, in our more general context, the precise-probabilistic notion of a representing probability measure on the simplex of all probability mass functions is replaced by a Bernstein coherent set of polynomials on this simplex. This set of polynomials serves completely the same purpose as the representing probability measure: it completely determines, and conveniently and densely summarises, all predictive inferences. This is the reason why the rest of the developments in the paper are expressed in terms of such Bernstein coherent sets of polynomials.

We introduce coherent inference systems in Section 6 as maps that associate with any finite set of categories a Bernstein coherent set of polynomials on the simplex of probability mass functions on that set. So a coherent inference system is a way of fixing completely all coherent predictive inferences for all possible category sets. Our reasons for introducing such coherent inference systems are twofold. First of all, the inference principles in Sec-

3. In contradistinction with de Finetti's version, our version has no problems with conditioning on observed sequences of (lower) probability zero.

tion 4 impose connections between predictive inferences for different category sets, so we can represent such inference principles mathematically as restrictions on coherent inference systems, which is the main topic of Section 7. Secondly, it allows us to extend the method of natural extension—or conservative inference—introduced in Section 2.2, to also take into account principles for predictive inference, or more generally, predictive inference for multiple category sets at once. This leads to a method of combining (local) predictive probability assessments with (global) inference principles to produce the most conservative predictive inferences compatible with them.

As a first illustration of the power of our methodology, we look at immediate prediction in Section 8: what implications do representation insensitivity and specificity have for predictive inference about the single next observation? We show that our approach allows us to streamline, simplify and significantly extend previous attempts in this direction by De Cooman et al. (2009a).

The material in Sections 9–14 shows, by producing explicit examples, that there are quite a few different types—even uncountable infinities—of coherent inference systems that are representation insensitive and/or specific. We discuss the vacuous and nearly vacuous inference systems in Sections 9 and 10, the skeptically cautious inference system in Section 11, the family of IDMM inference systems in Section 12, the family of skeptical IDMM inference systems in Section 13, and the Haldane inference system in Section 14. Most of these inference systems, apart from the IDMM, appear here for the first time. Also, we believe that we are the first to publish a detailed and explicit—as well as still elegant—proof that the IDMM inference systems are indeed representation insensitive and specific. It should already be mentioned here, however, that our IDMM inference systems are based on a modified, and arguably better behaved, version of the models originally introduced by Walley and Bernard (see Walley, 1996; Walley & Bernard, 1999; Bernard, 2005); we refer to Appendix D for more explanation, with a proof that the original IDMM is not specific and that, contrary to what is often claimed, it does not satisfy the so-called nestedness property.

Our results disprove the conjecture (Bernard, 2007; De Cooman et al., 2009a) that the IDMM inference systems—our version or the original one—are the only ones, or even the most conservative ones, that satisfy both representation insensitivity and specificity. But we do show in Section 15 that the IDMM family of immediate predictions—which are the same for our version and for the original one—are in a definite sense the most conservative ones that are representation insensitive and specific, and satisfy another requirement, which we have called 'having concave surprise'.

In the conclusion (Section 16) we point to a number of surprising consequences of our results, and discuss avenues for further research.

In order to make this paper as self-contained as possible, we have included a number of appendices with additional discussion. To help the reader find his way through the many notions and notations we need in this paper, Appendix A provides of list of the most common ones, with a short hint at their meaning, and where they are introduced. Appendix B provides useful and necessary background on the theory of multivariate polynomials on simplices, and the important part that Bernstein basis polynomials have there. Our discussion of IDMM inference systems relies quite heavily on Dirichlet densities on simplices, and the expectation operators associated with them. We discuss their most important and relevant properties in Appendix C. Appendix D contains a discussion of the original IDM and IDMM models, as proposed by Walley and Bernard (see Walley, 1991, 1996; Walley & Bernard, 1999; Bernard, 2005), where we show that some of the claims they make about this model need to be more carefully formulated. As we stated above, this is our main reason for introducing, in Section 12, our own modified version of the IDMM models, which does not suffer from such shortcomings, and produces the same immediate prediction models as the original version. Finally, in an effort to make this lengthy paper as readable as possible, we have moved all proofs, and some additional technical discussion, to Appendix E.

2. Imprecise Probability Models

In this section, we give a concise overview of imprecise probability models for representing, and making inferences and decisions under, uncertainty. As suggested in the Introduction, we shall focus on sets of desirable gambles as our uncertainty models of choice.

Let us briefly summarise in the next section why, in the present paper, we work with such sets as our basic uncertainty models for doing conservative probabilistic inference. The reader who wants to dispense with motivation can proceed to Section 2.2, where we introduce the mathematics behind these models. In later sections, we shall of course also briefly mention derived results in terms of the more familiar language of (lower) previsions and probabilities.

2.1 Why Sets of Desirable Gambles?

First of all, a number of examples in the literature (Moral, 2005; Couso & Moral, 2011; De Cooman & Quaeghebeur, 2012; De Cooman & Miranda, 2012) have shown that working with and making inferences using such models is more general and more expressive. It is also simpler and more elegant from a mathematical point of view, and it has a very intuitive geometrical interpretation (Quaeghebeur, 2014). We shall see in Sections 2.4 and 3 that marginalisation and conditioning are especially straightforward, and that there are no issues with conditioning on sets of (lower) probability zero.

Also, it should become apparent from the discussion in Section 2.2, and has been explained in some detail by Moral and Wilson (1995) and De Cooman and Miranda (2012), that the similarity between accepting a gamble on the one hand and accepting a proposition to be true on the other, gives a very 'logical' flavour to conservative probabilistic inference. Indeed, there is a strong analogy between the two, which connects conservative probabilistic inference—also called *natural extension* in the field—with logical deduction: where in classical propositional logic we are looking for the smallest deductively closed set that contains a number of given propositions, in an imprecise probabilities context we are looking for the smallest coherent set of desirable gambles that contains a number of given gambles. In the context of this analogy, precise probability models are closely related to complete, or maximal, deductively closed sets—perfect information states. This is a clear indication that precise probability models by themselves are not well suited for dealing with *conservative inference*, and that we need the broader context of imprecise probability models as a natural language and setting in which to do this. So in summary, working with sets of desirable gambles encompasses and subsumes as special cases both classical (or 'precise') probabilistic inference and inference in classical propositional logic; see the detailed discussion by De Cooman and Miranda (2012).

Finally, as we briefly explain in Section 5, De Cooman and Quaeghebeur (2012) have shown that working with sets of coherent desirable gambles is especially illuminating in the context of modelling exchangeability assessments: it exposes the simple geometrical meaning of the notion of exchangeability, and leads to a simple and particularly elegant proof of a significant generalisation of de Finetti's (1937) Representation Theorem for exchangeable random variables.

In summary, we work with sets of desirable gambles because they are the most powerful, expressive and general models at hand, because they are very intuitive to work with—though unfortunately less familiar to most people not closely involved in the field—, and, very importantly, because they avoid problems with conditioning on sets of (lower) probability zero. For more details, we refer to the work of Walley (2000), Moral (2005), Couso and Moral (2011), De Cooman and Quaeghebeur (2012), and Quaeghebeur (2014).

2.2 Coherent Sets of Desirable Gambles and Natural Extension

We consider a variable X that assumes values in some finite[4] possibility space A. We model a subject's beliefs about the value of X by looking at which gambles on this variable the subject finds *desirable*, meaning that he strictly prefers[5] them to the zero gamble—the status quo. This is a very general approach, that extends the usual rationalist and subjectivist approach to probabilistic modelling to allow for indecision and imprecision.

A *gamble* is a real-valued function f on A. It is interpreted as an uncertain reward $f(X)$ that depends on the value of X, and is expressed in units of some predetermined linear utility. It represents the reward the subject gets in a transaction where first the actual value x of X is determined, and then the subject receives the amount of utility $f(x)$—which may be negative, meaning he has to pay it. Throughout the paper, we use the device of writing $f(X)$ when we want to make clear what variable X the gamble f depends on.

Events are subsets of the possibility space A. With any event $B \subseteq A$ we can associate a special gamble $\mathbb{I}_B$, called its *indicator*, which assumes the value 1 on B and 0 elsewhere.

We denote the set of all gambles on A by $\mathcal{L}(A)$. It is a linear space under point-wise addition of gambles, and point-wise multiplication of gambles with real numbers. For any subset $\mathcal{A}$ of $\mathcal{L}(A)$, $\operatorname{posi}(\mathcal{A})$ is the set of all positive linear combinations of gambles in $\mathcal{A}$:

$$\operatorname{posi}(\mathcal{A}) \coloneqq \left\{ \sum_{k=1}^{n} \lambda_k f_k \colon f_k \in \mathcal{A},\, \lambda_k \in \mathbb{R}_{>0},\, n \in \mathbb{N} \right\}. \tag{1}$$

Here, $\mathbb{N}$ is the set of natural numbers (without zero), and $\mathbb{R}_{>0}$ is the set of all positive real numbers. A *convex cone* of gambles is a subset $\mathcal{A}$ of $\mathcal{L}(A)$ that is closed under positive linear combinations, meaning that $\operatorname{posi}(\mathcal{A}) = \mathcal{A}$.

4. For the sake of simplicity, we restrict this discussion to finite possibility spaces, because this is all we really need for the purposes of this paper. In a very limited number of remarks further on, we shall have occasion to mention related notions for infinite possibility spaces, but we will give ample references there to guide the interested reader to the relevant literature.
5. We want to point out that the notion of strict *preference*—or preference without indifference—commonly used in preference modelling, should not be confused with Walley's (1991, Section 3.7.7) notion of strict *desirability*, which is only one of the many ways to construct from a lower prevision a set of gambles that are strictly preferred to the zero gamble; see also the discussion near the end of Section 2.5. For more details, we refer to a recent paper by Quaeghebeur, De Cooman, and Hermans (2014).

For any two gambles f and g on A, we write '$f \geq g$' if $(\forall x \in A) f(x) \geq g(x)$, and '$f > g$' if $f \geq g$ and $f \neq g$. A gamble $f > 0$ is called *positive*. A gamble $g \leq 0$ is called *non-positive*. $\mathcal{L}_{>0}(A)$ denotes the convex cone of all positive gambles, and $\mathcal{L}_{\leq 0}(A)$ the convex cone of all non-positive gambles.

We collect the gambles that a subject finds desirable—strictly prefers[6] to the zero gamble—into his *set of desirable gambles*, and we shall take such sets as our basic uncertainty models. Of course, they have to satisfy certain rationality criteria:

Definition 1 (Coherence)**.** A set of desirable gambles $\mathcal{D} \subseteq \mathcal{L}(A)$ is called *coherent* if it satisfies the following requirements:

D1. $0 \notin \mathcal{D}$;

D2. $\mathcal{L}_{>0}(A) \subseteq \mathcal{D}$;

D3. $\mathcal{D} = \operatorname{posi}(\mathcal{D})$.

$\mathbb{D}(A)$ denotes the set of all coherent sets of desirable gambles on A.

Requirement D3 turns $\mathcal{D}$ into a *convex cone*. Due to D2, it includes $\mathcal{L}_{>0}(A)$; by D1–D3, it *avoids non-positivity*:

D4. if $f \leq 0$ then $f \notin \operatorname{posi}(\mathcal{D})$, or equivalently $\mathcal{L}_{\leq 0}(A) \cap \operatorname{posi}(\mathcal{D}) = \emptyset$.

$\mathcal{L}_{>0}(A)$ is the smallest coherent subset of $\mathcal{L}(A)$. This so-called *vacuous model* therefore reflects minimal commitments on the part of the subject: if he knows absolutely nothing about the likelihood of the different outcomes, he will only strictly prefer to zero those gambles that never decrease his wealth and have some possibility of increasing it.

When $\mathcal{D}_1 \subseteq \mathcal{D}_2$, a subject with a set of desirable gambles $\mathcal{D}_1$ is more conservative, or less committal, than a subject with a set of desirable gambles $\mathcal{D}_2$, simply because the latter strictly prefers to zero all the gambles that the former does, and possibly more. The inclusion relation imposes a natural partial ordering on sets of desirable gambles, with a simple interpretation of 'is at least as conservative as'.

For any non-empty family of coherent sets of desirable gambles $\mathcal{D}_i$, $i \in I$, its intersection $\bigcap_{i \in I} \mathcal{D}_i$ is still coherent. This simple result underlies the notion of (conservative) *coherent inference*. If a subject gives us an *assessment*—a set $\mathcal{A} \subseteq \mathcal{L}(A)$ of gambles on A that he finds desirable—then it tells us exactly when this assessment can be extended to a coherent set of desirable gambles, and how to construct the smallest—and therefore least committal or most conservative—such set:

Theorem 2 (Natural Extension, De Cooman & Quaeghebeur, 2012)**.** *Let $\mathcal{A} \subseteq \mathcal{L}(A)$, and define its* natural extension *by:*[7]

$$\mathcal{E}_{\mathcal{A}} := \bigcap \{\mathcal{D} \in \mathbb{D}(A) \colon \mathcal{A} \subseteq \mathcal{D}\}.$$

Then the following statements are equivalent:

6. See footnote 5.
7. As usual, in this expression, we let $\bigcap \emptyset = \mathcal{L}(A)$.

(i) $\mathcal{A}$ *avoids non-positivity:* $\mathcal{L}_{\leq 0}(A) \cap \operatorname{posi}(\mathcal{A}) = \emptyset$*;*

(ii) $\mathcal{A}$ *is included in some coherent set of desirable gambles;*

(iii) $\mathcal{E}_{\mathcal{A}} \neq \mathcal{L}(A)$*;*

(iv) *the set of desirable gambles* $\mathcal{E}_{\mathcal{A}}$ *is coherent;*

(v) $\mathcal{E}_{\mathcal{A}}$ *is the smallest coherent set of desirable gambles that includes* $\mathcal{A}$*.*

When any (and hence all) of these equivalent statements holds, $\mathcal{E}_{\mathcal{A}} = \operatorname{posi}(\mathcal{L}_{>0}(A) \cup \mathcal{A})$*. Moreover,* $\mathcal{A}$ *is coherent if and only if* $\mathcal{A} \neq \mathcal{L}(A)$ *and* $\mathcal{E}_{\mathcal{A}} = \mathcal{A}$*.*

2.3 Maximal Coherent Sets of Desirable Gambles

An element $\mathcal{D}$ of $\mathbb{D}(A)$ is called *maximal* if it is not strictly included in any other element of $\mathbb{D}(A)$, or in other words, if adding any gamble f to $\mathcal{D}$ makes sure we can no longer extend the set $\mathcal{D} \cup \{f\}$ to a set that is still coherent:

$$(\forall \mathcal{D}' \in \mathbb{D}(A))(\mathcal{D} \subseteq \mathcal{D}' \Rightarrow \mathcal{D} = \mathcal{D}').$$

$\mathbb{M}(A)$ denotes the set of all maximal elements of $\mathbb{D}(A)$. A coherent set of desirable gambles $\mathcal{D}$ is maximal if and only if for all non-zero gambles f on A, $f \notin \mathcal{D} \Rightarrow -f \in \mathcal{D}$ (see Couso & Moral, 2011 for the case of finite A, and De Cooman & Quaeghebeur, 2012 for the infinite case). Coherence and natural extension can be described completely in terms of maximal elements:

Theorem 3 (Couso & Moral, 2011; De Cooman & Quaeghebeur, 2012)**.** *A set* $\mathcal{A}$ *avoids non-positivity if and only if there is some maximal* $\mathcal{D} \in \mathbb{M}(A)$ *such that* $\mathcal{A} \subseteq \mathcal{D}$*. Moreover,* $\mathcal{E}_{\mathcal{A}} = \bigcap \{\mathcal{D} \in \mathbb{M}(A) \colon \mathcal{A} \subseteq \mathcal{D}\}$*.*

2.4 Conditioning with Sets of Desirable Gambles

Let us suppose that our subject has a coherent set $\mathcal{D}$ of desirable gambles on A, expressing his beliefs about the value that a variable X assumes in A. We can then ask what his so-called *updated* set $\mathcal{D}\rfloor B$ of desirable gambles on B would be, were he to receive the additional information—and nothing more—that X actually belongs to some subset B of A. The *updating*, or *conditioning*, *rule* for sets of desirable gambles states that:

$$g \in \mathcal{D}\rfloor B \Leftrightarrow g\mathbb{I}_B \in \mathcal{D} \text{ for all gambles } g \text{ on } B. \tag{2}$$

It states that the gamble g is desirable to a subject were he to observe that $X \in B$ if and only if the *called-off gamble* $g\mathbb{I}_B$ is desirable to him. This called-off gamble $g\mathbb{I}_B$ is the gamble on the variable X that gives a zero reward—is called off—unless $X \in B$, and in that case reduces to the gamble g on the new possibility space B. The updated set $\mathcal{D}\rfloor B$ is a set of desirable gambles on B that is still coherent, provided that $\mathcal{D}$ is (De Cooman & Quaeghebeur, 2012). See the discussions by Moral (2005), Couso and Moral (2011), De Cooman and Quaeghebeur (2012), De Cooman and Miranda (2012) and Quaeghebeur (2014) for more detailed information on updating sets of desirable gambles.

2.5 Coherent Lower Previsions

We now use coherent sets of desirable gambles to introduce derived concepts, such as coherent lower previsions, and probabilities.

Given a coherent set of desirable gambles $\mathcal{D}$, the functional $\underline{P}$ defined on $\mathcal{L}(A)$ by

$$\underline{P}(f) \coloneqq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \mathcal{D}\} \text{ for all } f \in \mathcal{L}(A), \tag{3}$$

is a *coherent lower prevision* (Walley, 1991, Thm. 3.8.1). This means that it is a lower envelope of the expectations associated with some set of probability mass functions,[8] or, equivalently, that it satisfies the following coherence properties (Walley, 1991, 2000; De Cooman & Quaeghebeur, 2012; Miranda & De Cooman, 2014; Troffaes & De Cooman, 2014):

P1. $\underline{P}(f) \geq \min f$ for all gambles f on A;

P2. $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$ for all gambles f, g on A;

P3. $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ for all gambles f on A and all real $\lambda \geq 0$.

Here we used the notation $\min f \coloneqq \min\{f(x) \colon x \in A\}$; $\max f$ is defined similarly. The *conjugate upper prevision* $\overline{P}$ is defined by $\overline{P}(f) \coloneqq \inf\{\mu \in \mathbb{R} \colon \mu - f \in \mathcal{D}\} = -\underline{P}(-f)$. The following properties are implied by P1–P3:

P4. $\max f \geq \overline{P}(f) \geq \underline{P}(f) \geq \min f$ for all gambles f on A;

P5. $\underline{P}(f+\mu) = \underline{P}(f) + \mu$ and $\overline{P}(f+\mu) = \overline{P}(f) + \mu$ for all gambles f on A and all $\mu \in \mathbb{R}$.

For any gamble f, $\underline{P}(f)$ is called the *lower prevision* of f, and it follows from Equation (3) that it can be interpreted as the subject's supremum desirable price for buying the gamble f. For any event B, $\underline{P}(\mathbb{I}_B)$ is also denoted by $\underline{P}(B)$, and called the *lower probability* of B; it can be interpreted as the subject's supremum desirable rate for betting on B. Similarly for upper previsions and upper probabilities.

The lower prevision associated with the vacuous set of desirable gambles $\mathcal{L}_{>0}(A)$ is given by $\underline{P}(f) = \min f$. It is called the *vacuous* lower prevision, and it is the point-wise smallest, or most conservative, of all coherent lower previsions.

The coherent conditional model $\mathcal{D}\rfloor B$, with B a non-empty subset of A, induces a *conditional lower prevision* $\underline{P}(\cdot|B)$ on $\mathcal{L}(B)$, by invoking Equation (3):

$$\underline{P}(g|B) \coloneqq \sup\{\mu \in \mathbb{R} \colon g - \mu \in \mathcal{D}\rfloor B\} = \sup\{\mu \in \mathbb{R} \colon [g-\mu]\mathbb{I}_B \in \mathcal{D}\} \quad \text{for all gambles } g \text{ on } B. \tag{4}$$

It is not difficult to show (Walley, 1991) that $\underline{P}$ and $\underline{P}(\cdot|B)$ are related through the following coherence condition:

$$\underline{P}([g - \underline{P}(g|B)]\mathbb{I}_B) = 0 \text{ for all } g \in \mathcal{L}(B), \tag{GBR}$$

8. This statement is valid because we are working with finite A. For infinite A, similar results can be shown to hold (Walley, 1991; De Cooman & Quaeghebeur, 2012; Miranda & De Cooman, 2014; Troffaes & De Cooman, 2014), and then the expectations involved are coherent previsions—expectation operators associated with finitely additive probability measures. See also the discussion in Section 2.6.

called the *Generalised Bayes Rule*. This rule allows us to infer $\underline{P}(\cdot|B)$ uniquely from $\underline{P}$, provided that $\underline{P}(B) > 0$. Otherwise, there is usually an infinity of coherent lower previsions $\underline{P}(\cdot|B)$ that are coherent with $\underline{P}$ in the sense that they satisfy (GBR), or equivalently, that there is some coherent set of desirable gambles $\mathcal{D}$ that leads to both $\underline{P}$ and $\underline{P}(\cdot|B)$. Two particular conditioning rules, namely *natural* and *regular extension* (Walley, 1991; Miranda & De Cooman, 2014), always produce conditional lower previsions that satisfy GBR, and are therefore coherent with $\underline{P}$. When $\overline{P}(B) > 0$—but not necessarily when $\overline{P}(B) = 0$!—they always produce the point-wise smallest and largest coherent conditional lower previsions, respectively (Miranda, 2009; Miranda & De Cooman, 2014).[9]

Many different coherent sets of desirable gambles lead to the same coherent lower prevision $\underline{P}$, and they typically differ only in their boundaries. In this sense, coherent sets of desirable gambles are more informative than coherent lower previsions: a gamble with positive lower prevision is always desirable and one with a negative lower prevision never, but a gamble with zero lower prevision lies on the border of the set of desirable gambles, and the lower prevision does not generally provide information about the desirability of such gambles. If such border behaviour is important—and it is when dealing with conditioning on events with zero (lower) probability (Walley, 2000; Moral, 2005; Couso & Moral, 2011; Quaeghebeur, 2014)—it is useful to work with sets of desirable gambles rather than lower previsions, because as Equations (2) and (4) tell us, they allow us to derive unique conditional models from unconditional ones: with a coherent set of desirable gambles $\mathcal{D}$ there corresponds a unique conditional set of desirable gambles $\mathcal{D}\rfloor B$ and a unique conditional lower prevision $\underline{P}(\cdot|B)$, for any non-empty event B. The smallest set of desirable gambles that induces a given coherent lower prevision, is called the associated set of *strictly desirable* gambles (Walley, 1991) and is given by $\{f \in \mathcal{L}(A) \colon f > 0 \text{ or } \underline{P}(f) > 0\}$. See the papers by Walley (2000) and Quaeghebeur (2014) for additional discussion about why sets of desirable gambles are more informative than coherent lower previsions.

2.6 Linear Previsions and Credal Sets

When the coherent lower and the upper prevision coincide on all gambles, then the real functional P defined on $\mathcal{L}(A)$ by $P(f) \coloneqq \underline{P}(f) = \overline{P}(f)$ for all $f \in \mathcal{L}(A)$ is a *coherent prevision*. Since we assumed that A is finite,[10] this means that it corresponds to the expectation operator associated with a probability mass function p: $P(f) = \sum_{x\in A} f(x)p(x) =: E_p(f)$ for all $f \in \mathcal{L}(A)$, where $p(x) \coloneqq P(\mathbb{I}_{\{x\}})$ for all $x \in A$. This happens in particular if the lower and upper previsions are induced by a maximal coherent set of desirable gambles. Indeed, up to boundary behaviour, the so-called *precise probability models* P correspond to maximal coherent sets of desirable gambles; see the discussions by Williams (1975a), Miranda and Zaffalon (2011, Proposition 6) and Couso and Moral (2011, Section 5) for more information.

9. The conditional lower previsions in Section 12 on the IDMM are produced by regular extension. The models in Sections 11, 13 and 14 have the same lower previsions amongst them, but in nearly all cases have very different conditional lower previsions, even though in these cases the natural and regular extensions coincide—they are vacuous there.
10. As already hinted at in footnote 8, similar things can still be said for infinite A, but this would unduly complicate the discussion. For more details, see the work by Walley (1991), Troffaes and De Cooman (2014) and Miranda and De Cooman (2014).

For coherent previsions P, the Generalised Bayes Rule (GBR) reduces to *Bayes's Rule*:

$$P(g\mathbb{I}_B) = P(B)P(g|B) \text{ for all } g \in \mathcal{L}(B), \tag{BR}$$

indicating that this central probabilistic updating rule is a special case of Equation (2).

Because we assumed that A is finite, we can define the so-called *credal set* $\mathcal{M}(\underline{P})$ associated with a coherent lower prevision $\underline{P}$ as:

$$\mathcal{M}(\underline{P}) \coloneqq \{p \in \Sigma_A \colon (\forall f \in \mathcal{L}(A)) E_p(f) \geq \underline{P}(f)\},$$

which is a closed and convex subset of the so-called simplex Σ_A of all probability mass functions on A.[11] Then $\underline{P}$ is the lower envelope of $\mathcal{M}(\underline{P})$: $\underline{P}(f) = \min\{E_p(f) \colon p \in \mathcal{M}(\underline{P})\}$ for all $f \in \mathcal{L}(A)$ (Walley, 1991; Miranda & De Cooman, 2014; Troffaes & De Cooman, 2014). In this sense, such convex closed sets of precise probability models can also be seen as imprecise probability models, and they are mathematically equivalent to coherent lower previsions. They are therefore also less general and powerful than coherent sets of desirable gambles, and also suffer from problems with conditioning on events with (lower) probability zero.[12]

3. Predictive Inference

Predictive inference, in the specific sense we are focussing on here, considers a number of variables $X_1, \dots, X_n$ assuming values in the same category set A—we define a *category set* as any non-empty *finite* set.[13] In what follows, we shall have occasion to use many different category sets, and we shall use italic capitals such as A, B, C, D, ... to refer to them.

We start our discussion of predictive inference models in the most general and representationally powerful language: coherent sets of desirable gambles, as introduced in the previous section. Further on, we shall also pay some attention to more specific derived models, such as predictive lower previsions, and predictive lower probabilities.

Predictive inference assumes generally that a number $\check{n}$ of observations have been made, so we know the values $\check{\boldsymbol{x}} = (x_1, \dots, x_{\check{n}})$ of the first $\check{n}$ variables $X_1, \dots, X_{\check{n}}$. Based on this *observation sample* $\check{\boldsymbol{x}}$, a subject then has a posterior *predictive model* $\mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}$ for the values that the next $\hat{n}$ variables $X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}$ assume in $A^{\hat{n}}$. This $\mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}$ is a coherent set of desirable gambles $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})$ on $A^{\hat{n}}$. Here we assume that $\hat{n} \in \mathbb{N}$. On the other hand, we want to allow that $\check{n} \in \mathbb{N}_0 \coloneqq \mathbb{N} \cup \{0\}$, which is the set of all natural numbers with zero: we also want to be able to deal with the case where no previous observations have

11. See Section 5.2 for an explicit definition of Σ_A.

12. Using sets of full conditional measures (Dubins, 1975; Cozman, 2013), rather than sets of probability mass functions, leads to an imprecise probability model that is related to sets of desirable gambles (Couso & Moral, 2011), and has no problems with conditioning on sets of lower probability zero either, but we feel it is less elegant and mathematically more complicated.

13. For formal reasons, we include the trivial case of category sets with a single element, in which case we are certain about the value that the variables assume.

been made. In that case, we call the corresponding model $\mathcal{D}_A^{\hat{n}}$ a *prior predictive model.*[14] Of course, technically speaking, $\check{n} + \hat{n} \leq n$.

As we said, the subject may also have a prior, unconditional model, for when no observations have yet been made. In its most general form, this will be a coherent set $\mathcal{D}_A^n$ of desirable gambles $f(X_1, \dots, X_n)$ on A^n, for some $n \in \mathbb{N}$. He may also have coherent sets $\mathcal{D}_A^{\hat{n}}$ of desirable gambles $f(X_1, \dots, X_{\hat{n}})$ on $A^{\hat{n}}$, where $\hat{n}$ can be any natural number such that $\hat{n} \leq n$; and the sets $\mathcal{D}_A^{\hat{n}}$ and $\mathcal{D}_A^n$ must then be related to each other through the following *marginalisation*, or *time consistency*, requirement:[15]

$$f(X_1, \dots, X_{\hat{n}}) \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f(X_1, \dots, X_{\hat{n}}) \in \mathcal{D}_A^n \text{ for all gambles } f \text{ on } A^{\hat{n}}. \tag{5}$$

In this expression, and throughout this paper, we identify a gamble f on $A^{\hat{n}}$ with its *cylindrical extension* f' on A^n, defined by $f'(x_1, \dots, x_{\hat{n}}, \dots, x_n) \coloneqq f(x_1, \dots, x_{\hat{n}})$ for all $(x_1, \dots, x_n) \in A^n$. If we introduce the marginalisation operator $\operatorname{marg}_{\hat{n}}(\cdot) \coloneqq \cdot \cap \mathcal{L}(A^{\hat{n}})$, then the time consistency condition can also be rewritten simply as $\mathcal{D}_A^{\hat{n}} = \operatorname{marg}_{\hat{n}}(\mathcal{D}_A^n) = \mathcal{D}_A^n \cap \mathcal{L}(A^{\hat{n}})$.

Prior (unconditional) predictive models $\mathcal{D}_A^n$ and posterior (conditional) ones $\mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}}$ must also be related through the following *updating* requirement:

$$f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}) \in \mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}} \Leftrightarrow f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}) \mathbb{I}_{\{\check{\boldsymbol{x}}\}}(X_1, \dots, X_{\check{n}}) \in \mathcal{D}_A^n \\ \text{for all gambles } f \text{ on } A^{\hat{n}}, \tag{6}$$

which is a special case of Equation (2): the gamble $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})$ is desirable after observing the sample $\check{\boldsymbol{x}}$ if and only if the gamble $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})\mathbb{I}_{\{\check{\boldsymbol{x}}\}}(X_1, \dots, X_{\check{n}})$ is desirable before any observations are made. This called-off gamble $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})\mathbb{I}_{\{\check{\boldsymbol{x}}\}}(X_1, \dots, X_{\check{n}})$ is the gamble that gives zero reward—is called off—unless the first $\check{n}$ observations are $\check{\boldsymbol{x}}$, and in that case reduces to the gamble $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})$ on the remaining variables $X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}$. The updating requirement is a generalisation of Bayes's Rule for updating, and in fact reduces to it when the sets of desirable gambles lead to (precise) probability mass functions, as described in Section 2.6 and proved in detail by Walley (2000) and also by De Cooman and Miranda (2012). But contrary to Bayes's Rule for probability mass functions, the updating rule (6) for coherent sets of desirable gambles clearly does not suffer from problems when the conditioning event has (lower) probability zero: it allows us to infer a unique conditional model from an unconditional one, regardless of the (lower or upper) probability of the conditioning event. We refer to the work of De Cooman and Miranda (2012) for detailed discussions of marginalisation and updating of sets of desirable gambles in a many-variable context.

14. So the terms 'posterior' and 'prior' in association with predictive models indicate whether or not previous observations have been made. But, in order to avoid the well-known issues with temporal coherence (Zaffalon & Miranda, 2013), we are assuming here that the prior and posterior models are based on a subject's beliefs before any observations have been made, so the posterior models refer to hypothetical future situations.

15. See also the related discussion of this notion by De Cooman and Miranda (2008b) and De Cooman and Quaeghebeur (2012); it should not be confused with the *temporal consistency* discussed by Goldstein (1983, 1985) and Zaffalon and Miranda (2013).

As explained in Section 2.5, we can use the relationship (3) to derive *prior* (unconditional) *predictive lower previsions* $\underline{P}^{\hat{n}}_A(\cdot)$ on $\mathcal{L}(A^{\hat{n}})$ from the prior set $\mathcal{D}^n_A$ through:

$$\underline{P}^{\hat{n}}_A(f) \coloneqq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \mathcal{D}^n_A\} \text{ for all gambles } f \text{ on } A^{\hat{n}} \text{ and all } 1 \leq \hat{n} \leq n,$$

and *posterior* (conditional) *predictive lower previsions* $\underline{P}^{\hat{n}}_A(\cdot|\check{\boldsymbol{x}})$ on $\mathcal{L}(A^{\hat{n}})$ from the posterior sets $\mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}$ through:

$$\underline{P}^{\hat{n}}_A(f|\check{\boldsymbol{x}}) \coloneqq \sup\left\{\mu \in \mathbb{R} \colon f - \mu \in \mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}\right\} \text{ for all gambles } f \text{ on } A^{\hat{n}}.$$

Further on, we shall also want to condition predictive lower previsions on the additional information that $(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}) \in B^{\hat{n}}$, for some proper subset B of A. Using the ideas in Sections 2.4 and 2.5, this leads for instance to the following lower prevision:

$$\underline{P}^{\hat{n}}_A(g|\check{\boldsymbol{x}}, B^{\hat{n}}) \coloneqq \sup\left\{\mu \in \mathbb{R} \colon [g - \mu]\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}\right\} \text{ for all gambles } g \text{ on } B^{\hat{n}}, \tag{7}$$

which is the lower prevision $\underline{P}^{\hat{n}}_A(\cdot|\check{\boldsymbol{x}})$ conditioned on the event $B^{\hat{n}}$.

4. Principles for Predictive Inference

So far, we have introduced coherence, marginalisation and updating as basic rationality requirements that prior and posterior predictive inference models must satisfy. But it could be envisaged that other requirements—other inference principles—can be imposed on our inference models. Because we want to show further on how to deal with such additional requirements in a theory for conservative predictive inference, we now discuss, by way of examples, a number of additional conditions, which have been suggested by a number of authors as reasonable properties of—or requirements for—predictive inference models. We want to stress here that by considering these requirements as examples, we do not want to defend using them in all circumstances, or mean to suggest that they are always reasonable or useful. They are what they are: inference principles that we might want to impose, and whose implications for conservative predictive inference we might therefore want to investigate.

4.1 Pooling Invariance

We first consider Walley's (1996) notion of representation invariance, which we prefer to call *pooling invariance*. Consider any set of categories A, and a partition $\mathcal{B}$ of A with non-empty partition classes. We can of course consider the partition $\mathcal{B}$ as a set of categories as well. Therefore, in order to streamline the discussion and notation, we shall henceforth denote it by B—as stated before, we want to use italic capitals for category sets. Each of its elements—some subset C of A—corresponds to a single new category, which consists of the original categories $x \in C$ being pooled—considered as one. Denote by $\rho(x)$ the unique element of the partition B that an original category $x \in A$ belongs to. This leads us to consider a surjective (onto) map ρ from A to B.

We say that a gamble g on A^n *does not differentiate between pooled categories* when:

$$g(\boldsymbol{x}) = g(\boldsymbol{y}) \text{ for all } \boldsymbol{x}, \boldsymbol{y} \in A^n \text{ such that } (\forall k \in \{1, \dots, n\})\rho(x_k) = \rho(y_k),$$

which means that there is some gamble f on B^n such that:

$$(\forall \boldsymbol{x} \in A^n) g(\boldsymbol{x}) = f(\rho(x_1), \dots, \rho(x_n)).$$

The idea underlying this formula—or requirement—is that with a sample $\boldsymbol{x} = (x_1, \dots, x_n) \in A^n$, there corresponds a sample $\rho\boldsymbol{x} \coloneqq (\rho(x_1), \dots, \rho(x_n)) \in B^n$ of pooled categories. Pooling invariance requires that for gambles $g = f \circ \rho$ that do not differentiate between pooled categories, it should make no difference whether we make predictive inferences using the set of original categories A, or using the set of pooled categories B. More formally, in terms of predictive lower previsions:

$$\underline{P}_A^{\hat{n}}(f \circ \rho) = \underline{P}_B^{\hat{n}}(f) \text{ and } \underline{P}_A^{\hat{n}}(f \circ \rho | \check{\boldsymbol{x}}) = \underline{P}_B^{\hat{n}}(f | \rho\check{\boldsymbol{x}})$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } B^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}},$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ and } f \circ \rho \in \mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \rfloor \rho\check{\boldsymbol{x}}$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } B^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}}.$$

Pooling invariance seems a reasonable principle to uphold in cases where the category set is not known in full detail. In that case it is useful to start from a limited set of broadly defined categories, and allow the creation of new ones, by pooling or splitting old categories as the observations proceed. In this context, recall Walley's (1996) example: if we have a closed bag containing coloured marbles, what is the probability of drawing a red marble from it? With no further information, our subject has no idea about the colours of the marbles in the bag, making it difficult to construct a suitable detailed category set for such an experiment. After a few draws from the bag, if the predictive inference model used respects pooling invariance, the inferences that are made about red marbles when he uses the category set $\{\text{red}, \text{yellow}, \text{blue}, \text{other}\}$ should be the same as those using the category set $\{\text{red}, \text{non-red}\}$, where all the colours different from red are pooled together into a single category. It appears that pooling invariance is a typically useful principle, for instance, in sampling species problems, when one wants to assess the prevalence of a given species in certain area.

There is a special case of pooling invariance, called *embedding invariance*,[16] which concentrates on the case without prior observations. In terms of lower previsions:

$$\underline{P}_A^{\hat{n}}(f \circ \rho) = \underline{P}_B^{\hat{n}}(f) \text{ for all } \hat{n} \in \mathbb{N} \text{ considered, and all gambles } f \text{ on } B^{\hat{n}},$$

or alternatively, and more generally, in terms of sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ for all } \hat{n} \in \mathbb{N} \text{ considered, and all gambles } f \text{ on } B^{\hat{n}}.$$

4.2 Renaming Invariance

Besides pooling invariance, we may also require *renaming invariance*: as long as no confusion can arise, it should not matter for a subject's predictive inferences what names, or labels, he gives to the different categories.

16. Walley calls the underlying requirement that the (lower) probability of an event A should not depend on the possibility space into which A is embedded, the *Embedding Principle* (Walley, 1991, Section 5.5.1).

This may seem too trivial to even mention, and as far as we know, it is always implicitly taken for granted in predictive inference. But it will be well to devote some attention to it here, in order to distinguish it from the category permutation invariance to be discussed shortly, with which it is easily confused if we do not pay proper attention. If we have a renaming bijection (a one-to-one and onto map) λ between a set of original categories A and a set of renamed categories C, where we clearly distinguish between the elements of A and those of C, then with a sample $\boldsymbol{x} = (x_1, \dots, x_n) \in A^n$ of original categories, there corresponds a sample of renamed categories $\lambda\boldsymbol{x} := (\lambda(x_1), \dots, \lambda(x_n))$. And with a gamble f on the set C^n of renamed samples, there corresponds a gamble $f \circ \lambda$ on the set A^n of original samples. Clearly, we can then require that it should make no difference whether we make predictive inferences using the set of original categories A, or using the set of renamed categories C. More formally, in terms of predictive lower previsions:

$$\underline{P}_A^{\hat{n}}(f \circ \lambda) = \underline{P}_C^{\hat{n}}(f) \text{ and } \underline{P}_A^{\hat{n}}(f \circ \lambda|\check{\boldsymbol{x}}) = \underline{P}_C^{\hat{n}}(f|\lambda\check{\boldsymbol{x}})$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } C^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}},$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \lambda \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_C^{\hat{n}} \text{ and } f \circ \lambda \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_C^{\hat{n}}\rfloor\lambda\check{\boldsymbol{x}}$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } C^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}}.$$

4.3 Category Permutation Invariance

We shall be especially interested in predictive inference where a subject starts from a state of prior ignorance. In such a state, he has no reason to distinguish between the different elements of any set of categories A he has chosen. To formalise this idea, consider a permutation ϖ of the elements of A.[17] With a sample $\boldsymbol{x}$ in A^n, there corresponds a permuted sample $\varpi\boldsymbol{x} := (\varpi(x_1), \dots, \varpi(x_n))$. And with any gamble f on A^n, there corresponds a permuted gamble $f \circ \varpi$ on A^n. If a subject has no reason to distinguish between categories z and their images $\varpi(z)$, it make sense to require the following *category permutation invariance*:[18]

$$\underline{P}_A^{\hat{n}}(f \circ \varpi) = \underline{P}_A^{\hat{n}}(f) \text{ and } \underline{P}_A^{\hat{n}}(f \circ \varpi|\check{\boldsymbol{x}}) = \underline{P}_A^{\hat{n}}(f|\varpi\check{\boldsymbol{x}})$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } A^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}},$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \varpi \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_A^{\hat{n}} \text{ and } f \circ \varpi \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_A^{\hat{n}}\rfloor\varpi\check{\boldsymbol{x}}$$
$$\text{for all } \check{n}, \hat{n} \in \mathbb{N} \text{ considered, all gambles } f \text{ on } A^{\hat{n}} \text{ and all } \check{\boldsymbol{x}} \in A^{\check{n}}.$$

Formally, this requirement closely resembles renaming invariance, but whereas the latter is a trivial requirement, category permutation invariance is a symmetry requirement between

17. This permutation ϖ of the elements of A, or in other words of the *categories*, should be contrasted with permutations π of the order of the observations, i.e. of the time set $\{1, \dots, n\}$, considered in our discussion of exchangeability, further on in Section 5.

18. This requirement is related to the notion of (weak) permutation invariance that De Cooman and Miranda (2007) have studied in much detail in a paper dealing with symmetry in uncertainty modelling. It goes back to Walley's (1991, Section 5.5.1) *Symmetry Principle*.

categories that can only be justified when our subject has no reason to distinguish between them, which may for instance be justified when he starts out from a state of prior ignorance. To draw attention to the difference between the two in a somewhat loose manner: category permutation invariance allows for confusion between new and old categories, something which renaming invariance carefully avoids.

To see why such a principle could be reasonable, recall Walley's (1996) bag of marbles example, introduced above when discussing pooling invariance. Since, before having drawn any marbles from the bag, our subject has no idea how the marbles are coloured, he is in a state of complete prior ignorance. Therefore, if he starts out with the sample space $\{\text{red}, \text{non-red}\}$, and observes the outcomes of a few draws, say twice non-red, he can consider the probability of obtaining a red marble on the next draw. But due to the symmetry originating in complete ignorance, if he were to permute the categories, calling the red marbles 'non-red' and the non-red ones 'red', the situation he is now looking at is completely the same as before, and therefore his probability of obtaining a non-red marble on the next draw after observing twice red, must be the same as that for observing a red one, after observing non-red twice. This principle is reminiscent of the Axiom A8 proposed by Carnap (1952) for his system of inductive logic. Of course, this is not a reasonable principle when our subject has some prior knowledge about the problem that would, for instance, allow him to impose an ordering on the categories.

4.4 Representation Insensitivity

We shall call *representation insensitivity* the combination of pooling, renaming and category permutation invariance. It means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are insensitive to the choice of representation—the category set. It is not difficult to see that representation insensitivity can be formally characterised as follows. Consider two category sets A and D such that there is a so-called *relabelling map* $\rho\colon A \to D$ that is *onto*, i.e. such that $D = \rho(A) \coloneqq \{\rho(x)\colon x \in A\}$. Then with a sample $\boldsymbol{x}$ in A^n, there corresponds a transformed sample $\rho\boldsymbol{x} \coloneqq (\rho(x_1), \dots, \rho(x_n))$ in D^n. And with any gamble f on D^n there corresponds a gamble $f \circ \rho$ on A^n.

4.4.1 Representation Insensitivity

For all category sets A and D such that there is an onto map $\rho\colon A \to D$, all $\check{n}, \hat{n} \in \mathbb{N}$ considered, all $\check{\boldsymbol{x}} \in A^{\check{n}}$ and all gambles f on $D^{\hat{n}}$:

$$\underline{P}_A^{\hat{n}}(f \circ \rho) = \underline{P}_D^{\hat{n}}(f) \text{ and } \underline{P}_A^{\hat{n}}(f \circ \rho|\check{\boldsymbol{x}}) = \underline{P}_D^{\hat{n}}(f|\rho\check{\boldsymbol{x}}), \tag{RI1}$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_D^{\hat{n}} \text{ and } f \circ \rho \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_D^{\hat{n}}\rfloor\rho\check{\boldsymbol{x}}. \tag{RI2}$$

There is also the weaker combination of pooling, renaming and category permutation invariance for models with no prior observations.

4.4.2 Prior Representation Insensitivity

For all category sets A and D such that there is an onto map $\rho\colon A \to D$, all $\hat{n} \in \mathbb{N}$ considered and all gambles f on $D^{\hat{n}}$:

$$\underline{P}^{\hat{n}}_{A}(f \circ \rho) = \underline{P}^{\hat{n}}_{D}(f), \tag{EI1}$$

or alternatively, and more generally, in terms of sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}^{\hat{n}}_{A} \Leftrightarrow f \in \mathcal{D}^{\hat{n}}_{D}. \tag{EI2}$$

4.5 Specificity

We now turn to another, rather peculiar but in our view intuitively appealing, potential property of predictive inferences. Assume that in addition to observing a sample of observations $\check{\boldsymbol{x}}$ of $\check{n}$ observations in a category set A, our subject comes to know or determine in some way that the $\hat{n}$ following observations will belong to a proper subset B of A, and nothing else—we might suppose for instance that an observation of $(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})$ has been made, but that it is imperfect, and only allows him to conclude that $(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}) \in B^{\hat{n}}$.

We can then impose the following requirement, which uses models conditioned on the event $B^{\hat{n}}$. Such conditional models have been introduced through Equations (2) and (4); see also the discussion leading to Equation (7), near the end of Section 3.

4.5.1 Specificity

For all category sets A and B such that $B \subseteq A$, all $\check{n}, \hat{n} \in \mathbb{N}$ considered, all $\check{\boldsymbol{x}} \in A^{\check{n}}$ and all gambles f on $B^{\hat{n}}$:

$$\underline{P}^{\hat{n}}_{A}(f|B^{\hat{n}}) = \underline{P}^{\hat{n}}_{B}(f) \text{ and } \underline{P}^{\hat{n}}_{A}(f|\check{\boldsymbol{x}}, B^{\hat{n}}) = \underline{P}^{\hat{n}}_{B}(f|\check{\boldsymbol{x}}{\downarrow_B}), \tag{SP1}$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}^{\hat{n}}_{A} \Leftrightarrow f \in \mathcal{D}^{\hat{n}}_{B} \text{ and } f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}^{\hat{n}}_{A}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}^{\hat{n}}_{B}\rfloor\check{\boldsymbol{x}}{\downarrow_B}, \tag{SP2}$$

where $\check{\boldsymbol{x}}{\downarrow_B}$ is the tuple of observations obtained by eliminating from the tuple $\check{\boldsymbol{x}}$ all observations not in B. In these expressions, when $\check{\boldsymbol{x}}{\downarrow_B}$ is the empty tuple, so when no observations in $\check{\boldsymbol{x}}$ are in B, the 'posterior' predictive model is simply taken to reduce to the 'prior' predictive model.

Specificity means that *the predictive inferences that a subject makes are the same as the ones he would get by focussing on the category set B, and at the same time discarding all the* previous *observations producing values outside B, in effect only retaining the observations that were inside B!* It is as if knowing that the future observations belong to B allows our subject to ignore all the previous observations that happened to lie outside B. The term specificity in this context seems to have been proposed by Bernard (1997, 2005), based on work by Rouanet and Lecoutre (1983). In a so-called specific inference approach, for questions, inferences and decisions involving only a restricted number of categories, a more general model can be replaced by a more specific model that deals only with the categories of interest, and if specificity is respected, the general and the specific models will produce the same inferences. Specificity seems to be a relevant principle when analysing categorical data that can be described by tree structures, as in the case of, for instance, patients that are classified according to symptoms (Bernard, 1997).

To give a very simple example involving, once again, Walley's bag of marbles, our subject may have observed, after some drawings, green, red, blue and white marbles. He is asked for his probability of drawing a red marble next, but some other observer has already seen what it is, and informs us that it is either green or red—perhaps due to bad lighting conditions or because she's colour blind. If the subject uses a specific inference model, he can disregard the previous observations involving other colours than green and red.

4.6 Prior Near-Ignorance

We use the notion of near-ignorance as defined by Walley (1991, p. 521) to give the following definition of prior near-ignorance in our context of predictive inference; see also the related discussions by Walley (1991, Section 5.3.2), Walley (1997, Section 3) and Walley and Bernard (1999, Section 2.3). We also refer to the paper by Piatti, Zaffalon, Trojani, and Hutter (2009) for an interesting discussion of why prior near-ignorance may produce undesirable results in certain contexts.

4.6.1 Prior Near-Ignorance

The prior model for any single variable X_k assuming values in some arbitrary category set A is vacuous, so for any category set A, any $\hat{n} \in \mathbb{N}$ considered, any $1 \leq k \leq \hat{n}$ and all gambles f on A:

$$\underline{P}_A^{\hat{n}}(\mathrm{ext}_k^{\hat{n}}(f)) = \min f,$$

or alternatively, and more generally, in terms of sets of desirable gambles:

$$\mathrm{ext}_k^{\hat{n}}(f) \in \mathcal{D}_A^{\hat{n}} \Rightarrow f > 0,$$

where $\mathrm{ext}_k^{\hat{n}}(f)$ denotes the *cylindrical extension* of f to a gamble on $A^{\hat{n}}$. It is defined by $\mathrm{ext}_k^{\hat{n}}(f)(x_1, \ldots, x_{\hat{n}}) \coloneqq f(x_k)$ for all $(x_1, \ldots, x_{\hat{n}}) \in A^{\hat{n}}$. A perhaps more intuitive, if less formally correct, notation for this gamble is $f(X_k)$.

Theorem 4. *Prior representation insensitivity implies prior near-ignorance.*

This simple result implies that no model all of whose predictive previsions are precise can be prior representation insensitive, let alone representation insensitive, as its prior model for immediate predictions should then be vacuous. We shall see in Section 14 that it is nevertheless possible for representation insensitive coherent inferences to deploy precise *posterior* predictive previsions.

5. Adding Exchangeability to the Picture

We are now, for the remainder of this paper, going to add two additional assumptions.

The *first assumption* is that there is, in principle, no upper bound on the number of variables that we can take into account. In other words, when we are considering n variables $X_1, \ldots, X_n$, we can always envisage looking at one more variable X_{n+1}. This effectively means that we are dealing with a *countably infinite sequence* of variables $X_1, \ldots, X_n, \ldots$ that assume values in the same category set A.

For our predictive inference models, this means that there is a sequence $\mathcal{D}_A^n$ of coherent sets of desirable gambles on A^n, $n \in \mathbb{N}$. This sequence should of course be *time-consistent* in the sense of Requirement (5), meaning that

$$(\forall n_1, n_2 \in \mathbb{N})(n_1 \leq n_2 \Rightarrow \mathcal{D}_A^{n_1} = \operatorname{marg}_{n_1}(\mathcal{D}_A^{n_2}) = \mathcal{D}_A^{n_2} \cap \mathcal{L}(A^{n_1})).$$

The *second assumption* is that this sequence of variables is *exchangeable*, which means, roughly speaking, that the subject believes that the order in which these variables are observed, or present themselves, has no influence on the decisions and inferences he will make regarding them.[19]

In this section, we explain succinctly how to deal with these assumptions technically, and what their consequences are for the predictive models we are interested in. For a detailed discussion and derivation of the results presented here, we refer to the papers by De Cooman et al. (2009b) and De Cooman and Quaeghebeur (2012).

We begin with some useful notation, which will be employed numerous times in what follows. Consider any element $\boldsymbol{\alpha} \in \mathbb{R}^A$. We consider $\boldsymbol{\alpha}$ as an *A-tuple*, with as many (real) components $\alpha_x \in \mathbb{R}$ as there are categories x in A. For any subset $B \subseteq A$, we then denote by $\alpha_B \coloneqq \sum_{x \in B} \alpha_x$ the sum of its components over B.

5.1 Permutations, Count Vectors and the Hypergeometric Distribution

Consider an arbitrary $n \in \mathbb{N}$. We denote by $\boldsymbol{x} = (x_1, \dots, x_n)$ a generic, arbitrary element of A^n. $\mathcal{P}^n$ is the set of all permutations π of the index set $\{1, \dots, n\}$. With any such permutation π, we can associate a permutation of A^n, also denoted by π, and defined by $(\pi\boldsymbol{x})_k \coloneqq x_{\pi(k)}$, or in other words, $\pi(x_1, \dots, x_n) \coloneqq (x_{\pi(1)}, \dots, x_{\pi(n)})$. Similarly, we lift π to a permutation π^t of $\mathcal{L}(A^n)$ by letting $\pi^t f \coloneqq f \circ \pi$, so $(\pi^t f)(\boldsymbol{x}) \coloneqq f(\pi\boldsymbol{x})$.

The permutation invariant atoms $[\boldsymbol{x}] \coloneqq \{\pi\boldsymbol{x} \colon \pi \in \mathcal{P}^n\}$, $\boldsymbol{x} \in A^n$ are the smallest permutation invariant subsets of A^n. We introduce the *counting map* $\boldsymbol{T} \colon A^n \to \mathcal{N}_A^n \colon \boldsymbol{x} \mapsto \boldsymbol{T}(\boldsymbol{x})$, where the *count vector* $\boldsymbol{T}(\boldsymbol{x})$ is the A-tuple with components

$$T_z(\boldsymbol{x}) \coloneqq |\{k \in \{1, \dots, n\} \colon x_k = z\}| \text{ for all } z \in A, \tag{8}$$

and the set of possible *count vectors* for n observations in A is given by

$$\mathcal{N}_A^n \coloneqq \left\{\boldsymbol{m} \in \mathbb{N}_0^A \colon m_A = n\right\}. \tag{9}$$

So $T_z(\boldsymbol{x})$ is the number of times the category z appears in the sample $\boldsymbol{x}$. If $\boldsymbol{m} = \boldsymbol{T}(\boldsymbol{x})$, then $[\boldsymbol{x}] = \{\boldsymbol{y} \in A^n \colon \boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{m}\}$, so the atom $[\boldsymbol{x}]$ is completely determined by the single count vector $\boldsymbol{m}$ of all its elements, and is therefore also denoted by $[\boldsymbol{m}]$.

We also consider the linear expectation operator $\mathrm{Hy}_A^n(\cdot|\boldsymbol{m})$ associated with the uniform distribution on the invariant atom $[\boldsymbol{m}]$:

$$\mathrm{Hy}_A^n(f|\boldsymbol{m}) \coloneqq \frac{1}{|[\boldsymbol{m}]|} \sum_{\boldsymbol{x} \in [\boldsymbol{m}]} f(\boldsymbol{x}) \text{ for all gambles } f \text{ on } A^n, \tag{10}$$

19. Exchangeability was also assumed by Carnap—his Axiom A7—and Johnson (1924), who named it the "permutation postulate".

where the number of elements $\nu(\boldsymbol{m}) \coloneqq |[\boldsymbol{m}]|$ in the invariant atom $[\boldsymbol{m}]$ is given by the *multinomial coefficient*:

$$\nu(\boldsymbol{m}) = \binom{m_A}{\boldsymbol{m}} = \binom{n}{\boldsymbol{m}} \coloneqq \frac{n!}{\prod_{z \in A} m_z!}. \tag{11}$$

This expectation operator in Equation (10) characterises—or is the one associated with—a (multivariate) *hyper-geometric distribution* (Johnson, Kotz, & Balakrishnan, 1997, Section 39.2), associated with random sampling without replacement from an urn with n balls of types $z \in A$, whose composition is characterised by the count vector $\boldsymbol{m}$. This is borne out by the fact that, for any $\boldsymbol{y} \in A^{n'}$, with $0 \leq n' \leq n$ and $\boldsymbol{m}' = \boldsymbol{T}(\boldsymbol{y})$,

$$\mathrm{Hy}_A^n(\mathbb{I}_{\{\boldsymbol{y}\}}|\boldsymbol{m}) = \begin{cases} \nu(\boldsymbol{m} - \boldsymbol{m}')/\nu(\boldsymbol{m}) & \text{if } \boldsymbol{m}' \leq \boldsymbol{m} \\ 0 & \text{otherwise} \end{cases}$$

is the probability of randomly selecting, without replacement, a sequence of n' balls of types $\boldsymbol{y}$ from an urn with n balls whose composition is determined by the count vector $\boldsymbol{m}$. See also the running example below for a more concrete illustration.

This hyper-geometric expectation operator can also be seen as a linear transformation Hy_A^n between the linear space $\mathcal{L}(A^n)$ and the generally much lower-dimensional linear space $\mathcal{L}(\mathcal{N}_A^n)$, turning a gamble f on A^n into a so-called *count gamble* $\mathrm{Hy}_A^n(f) \coloneqq \mathrm{Hy}_A^n(f|\cdot)$ on count vectors.

Running Example. In order to make our argumentation, and the notions we introduce and discuss, more tangible and concrete, we shall use a very simple running example, to which we shall come back repeatedly in a number of sections. The notations and assumptions made here will be maintained throughout the series.

Consider a (potentially infinite) sequence of coin flips, whose successive outcomes we denote by the variables $X_1, X_2, \dots X_n, \dots$ assuming values in the category set $\{H, T\}$. To make this somewhat more interesting than the usual run-of-the-mill example, assume that at each step—for each coin flip—Nathalie selects a coin from a bag of three coins, and hands it to Arthur, who then proceeds to flip it. The coin is then put back into the bag for the next step. The subject whose beliefs we are modelling, may or may not know something about the nature of the coins, or about how Nathalie is choosing the coins for the subsequent flips: she might choose them completely at random, or she might have a specific deterministic mechanism for selecting them, or ...

Consider the sequence $\check{\boldsymbol{x}} = (H, T, H, H)$ of the first $\check{n} = 4$ observed coin flips. The count vector $\boldsymbol{T}(\check{\boldsymbol{x}})$ that corresponds to this sequence is given by its components

$$T_H((H, T, H, H)) = 3 \text{ and } T_T((H, T, H, H)) = 1,$$

and we will denote it by $\check{\boldsymbol{m}} = (3, 1)$, letting the first component always refer to H, from now on. The corresponding permutation invariant atom is

$$[(H, T, H, H)] = [(3, 1)] = \{(T, H, H, H), (H, T, H, H), (H, H, T, H), (H, H, H, T)\}$$

and it has $\nu((3, 1)) = \frac{4!}{3!1!} = 4$ elements. The set of possible count vectors is given by $\mathcal{N}_{\{H,T\}}^4 = \{(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)\}$. Consider the event $\widehat{HT} \coloneqq \{(H, T), (T, H)\} \times$

$\{H, T\}^2$ of two different outcomes for the first two observations, then

$$\mathrm{Hy}^4_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \frac{1}{4}(1+1+0+0) = \frac{1}{2}$$

is the probability of observing two different outcomes in two random draws without replacement from an urn containing three balls marked H and one ball marked T, and whose composition is therefore determined by the count vector $(3,1)$. $\diamondsuit$

5.2 The Multinomial Distribution

Next, we consider the simplex Σ_A of all probability mass functions $\boldsymbol{\theta}$ on A:

$$\Sigma_A \coloneqq \left\{\boldsymbol{\theta} \in \mathbb{R}^A : \boldsymbol{\theta} \geq 0 \text{ and } \theta_A = 1\right\}, \text{ where, as before: } \theta_A \coloneqq \sum_{x \in A} \theta_x. \tag{12}$$

With a probability mass function $\boldsymbol{\theta} \in \Sigma_A$ on A, there corresponds the following *multinomial expectation* operator $\mathrm{Mn}^n_A(\cdot|\boldsymbol{\theta})$:[20]

$$\mathrm{Mn}^n_A(f|\boldsymbol{\theta}) \coloneqq \sum_{\boldsymbol{x} \in A^n} f(\boldsymbol{x}) \prod_{z \in A} \theta_z^{T_z(\boldsymbol{x})} \text{ for all gambles } f \text{ on } A^n, \tag{13}$$

which characterises the multinomial distribution, associated with n independent trials of an experiment with possible outcomes in A and probability mass function $\boldsymbol{\theta}$. Observe that

$$\begin{aligned}
\mathrm{Mn}^n_A(f|\boldsymbol{\theta}) &= \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} \left(\frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x} \in [\boldsymbol{m}]} f(\boldsymbol{x})\right) \nu(\boldsymbol{m}) \prod_{z \in A} \theta_z^{m_z} \\
&= \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} \mathrm{Hy}^n_A(f|\boldsymbol{m}) \nu(\boldsymbol{m}) \prod_{z \in A} \theta_z^{m_z} = \mathrm{CoMn}^n_A(\mathrm{Hy}^n_A(f)|\boldsymbol{\theta}),
\end{aligned}$$

where we used the so-called *count multinomial expectation* operator:[21]

$$\mathrm{CoMn}^n_A(g|\boldsymbol{\theta}) \coloneqq \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} g(\boldsymbol{m}) \nu(\boldsymbol{m}) \prod_{z \in A} \theta_z^{m_z} \text{ for all gambles } g \text{ on } \mathcal{N}^n_A. \tag{14}$$

Running Example. Consider $n = 4$ independent trials of an experiment with possible outcomes in the category set $\{H, T\}$ and probability mass function $\boldsymbol{\theta} = (\theta_H, \theta_T)$. Then

$$\mathrm{Mn}^4_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(\theta_H, \theta_T)) = 2\theta_H\theta_T^3 + 4\theta_H^2\theta_T^2 + 2\theta_H^3\theta_T = 2\theta_H\theta_T(\theta_H + \theta_T)^2 = 2\theta_H\theta_T,$$

gives the probability of the event $\widehat{HT}$. Observe, by the way, that $\mathrm{Mn}^n_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(\theta_H, \theta_T)) = 2\theta_H\theta_T$ for all $n \geq 2$.

With the gamble $f_{\widehat{HT}} \coloneqq \mathbb{I}_{\widehat{HT}}$ on observation sequences $(X_1, \dots, X_4)$, there corresponds a count gamble $g_{\widehat{HT}} \coloneqq \mathrm{Hy}^4_{\{H,T\}}(f_{\widehat{HT}}|\cdot)$ given by:

$$g_{\widehat{HT}}(0,4) = 0 \text{ and } g_{\widehat{HT}}(1,3) = \frac{1}{2} \text{ and } g_{\widehat{HT}}(2,2) = \frac{2}{3} \text{ and } g_{\widehat{HT}}(3,1) = \frac{1}{2} \text{ and } g_{\widehat{HT}}(4,0) = 0,$$

20. To avoid confusion, we make a (perhaps non-standard) distinction between the multinomial expectation, which is associated with sequences of observations, and the count multinomial expectation, associated with their count vectors.
21. See footnote 20.

and

$$\mathrm{CoMn}^4_{\{H,T\}}(g|(\theta_H,\theta_T)) = \frac{1}{2}4\theta_H\theta_T^3 + \frac{2}{3}6\theta_H^2\theta_T^2 + \frac{1}{2}4\theta_H^3\theta_T = 2\theta_H\theta_T$$

leads to the same polynomial as before, as it should. ◊

5.3 Multivariate Polynomials

Let us introduce the notation $\mathcal{N}_A \coloneqq \bigcup_{m\in\mathbb{N}} \mathcal{N}_A^m$ for the set of all possible count vectors corresponding to samples of at least one observation. In Equation (9), we can also let $n = 0$, which turns $\mathcal{N}_A^0$ into the singleton containing only the null count vector $\mathbf{0}$, all of whose components are zero. Then $\bigcup_{m\in\mathbb{N}_0} \mathcal{N}_A^m = \mathcal{N}_A \cup \{\mathbf{0}\}$ is the set of all possible count vectors.

For any such count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$, we consider the (multivariate) *Bernstein basis polynomial* $\mathrm{B}_{A,\boldsymbol{m}}$ of degree m_A on Σ_A, defined by:

$$\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \coloneqq \nu(\boldsymbol{m}) \prod_{z\in A} \theta_z^{m_z} = \binom{m_A}{\boldsymbol{m}} \prod_{z\in A} \theta_z^{m_z} \text{ for all } \boldsymbol{\theta} \in \Sigma_A. \tag{15}$$

In particular, of course, $\mathrm{B}_{A,\mathbf{0}} = 1$.

Any linear combination p of Bernstein basis polynomials of degree $n \geq 0$ is a (multivariate) polynomial on Σ_A, whose degree $\deg(p)$ is at most n.[22] We denote the linear space of all these polynomials of degree up to n by $\mathcal{V}^n(A)$. Of course, polynomials of degree zero are simply real constants. We have gathered relevant and useful information about multivariate polynomials in Appendix B. It follows from the discussion there that, for any $n \geq 0$, we can introduce a linear isomorphism CoMn_A^n between the linear spaces $\mathcal{L}(\mathcal{N}_A^n)$ and $\mathcal{V}^n(A)$: with any gamble g on $\mathcal{N}_A^n$, there corresponds a polynomial $\mathrm{CoMn}_A^n(g) \coloneqq \mathrm{CoMn}_A^n(g|\cdot) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} g(\boldsymbol{m})\mathrm{B}_{A,\boldsymbol{m}}$ in $\mathcal{V}^n(A)$, and conversely, for any polynomial $p \in \mathcal{V}^n(A)$ there is a unique gamble b_p^n on $\mathcal{N}_A^n$ such that $p = \mathrm{CoMn}_A^n(b_p^n)$.[23] Observe that in particular, for any $n \geq 0$ and $\boldsymbol{m} \in \mathcal{N}_A^n$:

$$\mathrm{CoMn}_A^n(\{\boldsymbol{m}\}|\boldsymbol{\theta}) = \mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Sigma_A. \tag{16}$$

We denote by $\mathcal{V}(A) \coloneqq \bigcup_{n\in\mathbb{N}_0} \mathcal{V}^n(A)$ the linear space of all (multivariate) polynomials on Σ_A, of arbitrary degree.

A set $\mathcal{H}_A \subseteq \mathcal{V}(A)$ of polynomials on Σ_A is called *Bernstein coherent* if it satisfies the following properties:

B1. $0 \notin \mathcal{H}_A$;

B2. $\mathcal{V}^+(A) \subseteq \mathcal{H}_A$;

B3. $\mathrm{posi}(\mathcal{H}_A) = \mathcal{H}_A$.

22. The degree may be smaller than n because the sum of all Bernstein basis polynomials of fixed degree is one. Strictly speaking, these polynomials p are restrictions to Σ_A of multivariate polynomials q on $\mathbb{R}^A$, called *representations* of p. For any p, there are multiple representations, with possibly different degrees. The smallest such degree is then called the degree $\deg(p)$ of p.

23. Strictly speaking, Equation (14) only defines the count multinomial expectation operator CoMn_A^n for $n > 0$, but it is clear that the definition extends trivially to the case $n = 0$.

Here, $\mathcal{V}^+(A)$ is the set of *Bernstein positive* polynomials on Σ_A: those polynomials p for which there is some $n \geq \deg(p)$ such that $b_p^n > 0$. It follows from Proposition 28 in Appendix B that $\mathcal{V}^+(A)$ is a subset of the set $\mathcal{V}^{++}(A)$ of all polynomials p such that $p(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$ in the interior $\operatorname{int}(\Sigma_A) := \{\boldsymbol{\theta} \in \Sigma_A : (\forall x \in A)\theta_x > 0\}$ of Σ_A. As a consequence of B1–B3, we find for the set $\mathcal{V}_0^-(A) := -\mathcal{V}^+(A)$ of *Bernstein negative* polynomials that:

B4. $\mathcal{V}_0^-(A) \cap \mathcal{H}_A = \emptyset$.

Finally, every Bernstein coherent set $\mathcal{H}_A$ of polynomials on Σ_A induces a lower prevision $\underline{H}_A$ on $\mathcal{V}(A)$ defined by:

$$\underline{H}_A(p) := \sup\{\mu \in \mathbb{R} : p - \mu \in \mathcal{H}_A\} \text{ for all } p \in \mathcal{V}(A). \tag{17}$$

This lower prevision is coherent, in the mathematical sense that it satisfies the coherence requirements P1–P3.[24]

5.4 Exchangeability and the Representation Theorem

We are now ready to deal with exchangeability. We shall give a definition for coherent sets of desirable gambles that generalises de Finetti's (1937, 1975) definition, and which allows for a significant generalisation of his Representation Theorem.

First of all, fix $n \in \mathbb{N}$. Then the subject considers the variables $X_1, \dots, X_n$ to be exchangeable when he does not distinguish between any gamble f on A^n and its permuted version $\pi^t f$, or in other words, if the gamble $f - \pi^t f$ is equivalent to the zero gamble for—or *indifferent* to—him. This means that he has a so-called *set of indifferent gambles*:

$$\mathcal{I}_A^n := \left\{f - \pi^t f : f \in \mathcal{L}(A^n) \text{ and } \pi \in \mathcal{P}^n\right\}.$$

If the subject also has a coherent set of desirable gambles $\mathcal{D}_A^n$, then this set must be compatible with the set of indifferent gambles $\mathcal{I}_A^n$, in the sense that it must satisfy the rationality requirement $\mathcal{D}_A^n + \mathcal{I}_A^n = \mathcal{D}_A^n$; see the detailed explanations and justifications by De Cooman and Quaeghebeur (2012) and Quaeghebeur et al. (2014) of this so-called *desiring sweetened deals* requirement. We then say that the sequence $X_1, \dots, X_n$, and the model $\mathcal{D}_A^n$, are *exchangeable*.

Next, the countably infinite sequence of variables $X_1, \dots, X_n \dots$ is called exchangeable if all the finite subsequences $X_1, \dots, X_n$ are, for $n \in \mathbb{N}$. This means that all models $\mathcal{D}_A^n$, $n \in \mathbb{N}$ are exchangeable. They should of course also be time-consistent.

We can now formulate a powerful generalisation of de Finetti's (1937, 1975) Representation Theorem, which is a straightforward compilation of various results proved by De Cooman and Quaeghebeur (2012):

Theorem 5 (Representation Theorem, De Cooman & Quaeghebeur, 2012)**.** *The sequence of sets $\mathcal{D}_A^n$ of desirable gambles on A^n, $n \in \mathbb{N}$ is coherent, time-consistent and exchangeable*

24. Actually, a suitably adapted version, where the underlying possibility space need no longer be finite (Walley, 1991; Troffaes & De Cooman, 2014), and where the domain is restricted to the polynomials on Σ_A (De Cooman & Quaeghebeur, 2012).

if and only if there is a Bernstein coherent set $\mathcal{H}_A$ of polynomials on Σ_A such that for all $\hat{n} \in \mathbb{N}$, all gambles f on $A^{\hat{n}}$, all $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and all $\check{\boldsymbol{x}} \in [\check{\boldsymbol{m}}]$:

$$f \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow \mathrm{Mn}_A^{\hat{n}}(f) \in \mathcal{H}_A \text{ and } f \in \mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}} \Leftrightarrow \mathrm{Mn}_A^{\hat{n}}(f) \mathrm{B}_{A,\check{\boldsymbol{m}}} \in \mathcal{H}_A. \tag{18}$$

In that case this representation *$\mathcal{H}_A$ is unique and given by $\mathcal{H}_A \coloneqq \bigcup_{n \in \mathbb{N}} \mathrm{Mn}_A^n(\mathcal{D}_A^n)$.*

It follows from Condition (18) that $\mathcal{H}_A$ *completely determines* all predictive inferences about the sequence of variables $X_1, \ldots, X_n, \ldots$, as it fixes all prior predictive models $\mathcal{D}_A^{\hat{n}}$ and all posterior predictive models $\mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}}$.[25] This tells us that the representation $\mathcal{H}_A$ is a set of polynomials that plays the same role as a probability measure, or density, or distribution function, on Σ_A in the precise-probabilistic case.

Indeed, the corresponding coherent lower prevision $\underline{H}_A$ on $\mathcal{V}(A)$ is given by Equation (17), and it can be shown to determine a convex closed (compact) set

$$\mathfrak{M}(\underline{H}_A) \coloneqq \{H_A \colon (\forall p \in \mathcal{V}(A)) H_A(p) \geq \underline{H}_A(p)\}$$

of coherent previsions H_A on $\mathcal{V}(A)$ (Walley, 1991; De Cooman et al., 2009b; De Cooman & Quaeghebeur, 2012; Troffaes & De Cooman, 2014). As we pointed out in footnote 2—and will come back to further on in footnote 36—each such coherent prevision H_A uniquely determines a σ-additive probability measure on the Borel sets of Σ_A, and therefore the set of polynomials $\mathcal{H}_A$, via $\mathfrak{M}(\underline{H}_A)$, uniquely determines a set of such probability measures. But, as we have argued before, $\mathcal{H}_A$ is *more informative* than $\underline{H}_A$ and $\mathfrak{M}(\underline{H}_A)$, and has no problems with conditioning on sets of lower probability zero: a Bernstein coherent set of polynomials $\mathcal{H}_A$ determines a unique lower prevision $\underline{H}_A$, and therefore through $\mathfrak{M}(\underline{H}_A)$ a unique set of probability measures—and densities if they are absolutely continuous—on the simplex Σ_A, but the converse is not necessarily—and usually not—the case. A set of probability densities can be used to define a coherent set of polynomials—we provide an example of how to do this in Section 12—but there will generally be more than one coherent set of polynomials that leads to this same set of densities, and the updating behaviour for these different sets of polynomials can be different on conditioning events of lower probability zero.

Condition (18) also tells us that the posterior predictive models $\mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}}$ only depend on the observed sequence $\check{\boldsymbol{x}}$ through the count vector $\check{\boldsymbol{m}} = \boldsymbol{T}(\check{\boldsymbol{x}})$: count vectors are *sufficient statistics* under exchangeability. For this reason, we shall from now on denote these posterior predictive models by $\mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{m}}$ as well as by $\mathcal{D}_A^{\hat{n}} \rfloor \check{\boldsymbol{x}}$. Also, every now and then, we shall use $\mathcal{D}_A^{\hat{n}} \rfloor \boldsymbol{0}$ as an alternative notation for $\mathcal{D}_A^{\hat{n}}$.

An immediate but interesting consequence of Theorem 5 is that updating on observations preserves exchangeability: after observing the values of the first $\check{n}$ variables, with count vector $\check{\boldsymbol{m}}$, the remaining sequence of variables $X_{\check{n}+1}, X_{\check{n}+2}, \ldots$ is still exchangeable, and Condition (18) tells us that its representation is given by the Bernstein coherent set of polynomials $\mathcal{H}_A \rfloor \check{\boldsymbol{m}}$ defined by:

$$\mathcal{H}_A \rfloor \check{\boldsymbol{m}} \coloneqq \{p \in \mathcal{V}(A) \colon \mathrm{B}_{A,\check{\boldsymbol{m}}} p \in \mathcal{H}_A\}. \tag{19}$$

25. This should be contrasted with the usual precise-probabilistic version, where the posterior predictive models are only uniquely determined if the observed sequences has non-zero probability; see also footnote 3.

If we compare this with Expressions (2) and (6), this tells us that, essentially, Bernstein basis polynomials serve as likelihood functions for updating sets of polynomials. We use $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ to refer to the coherent lower prevision on $\mathcal{V}(A)$ derived from $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ by means of Equation (17). For the special case $\check{\boldsymbol{m}} = \mathbf{0}$, we find that $\mathcal{H}_A\rfloor\mathbf{0} = \mathcal{H}_A$ and $\underline{H}_A(\cdot|\mathbf{0}) = \underline{H}_A$. Observe that $\underline{H}_A$ and $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ are related through the following version of the Generalised Bayes Rule:

$$\underline{H}_A([p - \underline{H}_A(p|\check{\boldsymbol{m}})]\mathrm{B}_{A,\check{\boldsymbol{m}}}) = 0 \text{ for all } p \in \mathcal{V}(A). \tag{20}$$

Clearly, $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ is completely determined by $\mathcal{H}_A$. One can consider $\mathcal{H}_A$ as a prior model on the parameter space Σ_A, and $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ plays the role of the posterior that is derived from it. We see from Condition (18) and Equation (19) that—similarly to what happens in a precise-probabilistic setting—the *multinomial distribution* serves as a direct link between on the one hand, the 'prior' $\mathcal{H}_A$ and its prior predictive inference models $\mathcal{D}_A^{\hat{n}}$ and, on the other hand, the 'posterior' $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ and its posterior predictive inference models $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}}$. Recalling our convention for $\check{\boldsymbol{m}} = \mathbf{0}$, we can summarise this as follows: for all $\hat{n} \in \mathbb{N}$ and all $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A^{\hat{n}})\colon\ \mathrm{Mn}_A^{\hat{n}}(f) \in \mathcal{H}_A\rfloor\check{\boldsymbol{m}}\right\} \tag{21}$$

and, as an immediate consequence:

$$\underline{P}_A^{\hat{n}}(f|\check{\boldsymbol{m}}) = \sup\left\{\mu \in \mathbb{R}\colon\ \mathrm{Mn}_A^{\hat{n}}(f) - \mu \in \mathcal{H}_A\rfloor\check{\boldsymbol{m}}\right\} \text{ for all } f \in \mathcal{L}(A^{\hat{n}}) \tag{22}$$

or, equivalently:

$$\underline{P}_A^{\hat{n}}(f|\check{\boldsymbol{m}}) = \underline{H}_A(\mathrm{Mn}_A^{\hat{n}}(f)|\check{\boldsymbol{m}}) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}). \tag{23}$$

From a practical point of view, Equation (23) will often be easier to work with than Equation (22), because as we shall see further on, $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ will often admit a simpler expression than $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$; compare Equations (45), (54), (61) and (69) with Equations (49), (55), (65) and (73), respectively. But, $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ is not always uniquely determined by $\underline{H}_A$: the relation (20) only allows us to determine $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ uniquely from $\underline{H}_A$ if the prior lower probability $\underline{H}_A(\mathrm{B}_{A,\check{\boldsymbol{m}}})$ of observing $\check{\boldsymbol{m}}$ is non-zero. Therefore, the sets of polynomials $\mathcal{H}_A$ are the more fundamental models, as they allow us to determine the $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ uniquely. As a quite dramatic illustration of this, we shall further on in Sections 11, 13 and 14 come across a number of quite different inference systems—with different $\mathcal{H}_A$—that give rise to the *same* prior $\underline{H}_A$ but different posterior $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$!

Running Example. We now assume that our subject assesses the sequence of coin flips to be exchangeable, and that he finds desirable any gamble of the type $\alpha - \mathbb{I}_{\{H\}}(X_n)$, for some fixed $\alpha \in (0,1]$; so his upper probability for observing heads on any coin flip is at most α. Since we infer from Equation (13) that for any $N \geq n$, $\mathrm{Mn}_{\{H,T\}}^N(\mathbb{I}_{\{H\}}(X_n)|\boldsymbol{\theta}) = \theta_H$, we infer from Theorem 5 that this assessment corresponds to the following coherent set of polynomials:

$$\mathcal{H}_\alpha \coloneqq \left\{\lambda_1 p^+ + \lambda_2(\alpha - \theta_H)\colon p^+ \in \mathcal{V}^+(\{H,T\}), \lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0} \text{ and } \max\{\lambda_1, \lambda_2\} > 0\right\},$$

which is the smallest Bernstein coherent set of polynomials that contains the polynomial $\alpha - \theta_H$; for more explanation, see also the discussions by De Cooman et al. (2009b) and De

Cooman and Quaeghebeur (2012). It then follows—after some manipulations—from Equation (17) and Proposition 28 that the corresponding lower prevision on $\mathcal{V}(\{H,T\})$ is completely determined by the following optimisation:

$$\underline{H}_{\alpha}(p)=\sup_{\lambda\geq 0}\min_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}}[p(\boldsymbol{\theta})+\lambda(\theta_H-\alpha)]$$

Hence, the lower probability of the event $\widehat{HT}$ is given by

$$\underline{H}_{\alpha}(2\theta_H\theta_T)=\sup_{\lambda\geq 0}\min_{x\in[0,1]}[2x(1-x)+\lambda(x-\alpha)]=0,$$

and its upper probability by

$$\begin{aligned}\overline{H}_{\alpha}(2\theta_H\theta_T)=-\underline{H}_{\alpha}(-2\theta_H\theta_T)&=\inf_{\lambda\geq 0}\max_{x\in[0,1]}[2x(1-x)-\lambda(x-\alpha)]\\&=\begin{cases}2\alpha(1-\alpha)&\text{if }\alpha\leq\frac{1}{2}\\\frac{1}{2}&\text{otherwise.}\end{cases}\end{aligned}$$

This tells us that exchangeability alone already guarantees that the upper probability of $\widehat{HT}$ is at most $\frac{1}{2}$. If all three coins in the bag are assumed to be biased towards heads, so $\alpha<\frac{1}{2}$, this upper probability drops below $\frac{1}{2}$. ♢

To finish this section on representation, we want to stress that the polynomials on Σ_A should not be given a behavioural interpretation as gambles that may or may not be desirable: they are merely mathematical and representational tools that help us characterise which gambles on observation sequences are desirable.[26] Similarly, the set of polynomials $\mathcal{H}_A$ and the lower prevision $\underline{H}_A$ are merely mathematical tools that allow for a more convenient representation of predictive models on observation sequences.

Running Example. To illustrate why the polynomial representation is so much more convenient and efficient, recall that if we want to make inferences about a sequence of coin flips of length n, we need to work with sets of desirable gambles on $\{H,T\}^n$, or in other words, with cones in a 2^n-dimensional space. If we work with their polynomial representations, we are led to consider cones of polynomials of degree up to n, which constitute a linear space that is spanned by the $n+1$ Bernstein basis polynomials of degree n, and is therefore only $n+1$-dimensional. Working with these polynomial representations therefore leads to a dramatic—exponential—reduction in complexity. ♢

6. Reasoning about Inference Systems

We have seen in the previous section that, once we fix a category set A, predictive inferences about exchangeable sequences assuming values in A are completely determined by a Bernstein coherent set $\mathcal{H}_A$ of polynomials on Σ_A. So if we had some way of associating a Bernstein coherent set $\mathcal{H}_A$ with every possible set of categories A, this would completely fix all predictive inferences. This leads us to the following definition.

26. It makes no operational, behavioural sense to consider the notion of 'accepting a polynomial', or finding it 'desirable'. This is very much like the classical case, where for de Finetti (1975) the probability distributions on the simplex Σ_A are only to be used as as mathematical representations, but have no direct behavioural meaning—although some Bayesians less careful about foundations than de Finetti might not care to make this distinction.

Definition 6 (Inference Systems)**.** We denote by $\mathbb{F}$ the collection of all *category sets*, i.e. finite non-empty sets. An *inference system* is a map Φ that maps any category set $A \in \mathbb{F}$ to some set of polynomials $\Phi(A) = \mathcal{H}_A$ on Σ_A. An inference system Φ is called *coherent* if for all category sets $A \in \mathbb{F}$, $\Phi(A)$ is a Bernstein coherent set of polynomials on Σ_A.

So, a coherent inference system is a way to systematically associate coherent predictive inferences with any category set. Since the inference principles in Section 4 impose connections between predictive inferences for different category sets, we now see that we can interpret these inference principles—or rather, represent them mathematically—as properties of, or restrictions on, coherent inference systems. This is what we shall do in Section 7, and it provides one important motivation for our introducing such systems. Another, equally important reason for doing so, is that it allows us to extend the method of natural extension—or conservative inference—introduced in Section 2.2, to also take into account inference principles for predictive inference, or more generally, predictive inference for multiple category sets at once.

To see how this comes about, let us show how we can do conservative reasoning with inference systems. For any two inference systems Φ_1 and Φ_2, we say that Φ_1 is *less committal*—or *more conservative*—than Φ_2, and we write $\Phi_1 \sqsubseteq \Phi_2$ if

$$(\forall A \in \mathbb{F})\Phi_1(A) \subseteq \Phi_2(A).$$

This simply means that the predictive inferences for each category set A are less committal for the first than for the second inference system. If we denote by $\mathbb{S}$ the set of all inference systems, then clearly this set is partially ordered by $\sqsubseteq$. Actually, it is a complete lattice, where the infimum and supremum of any non-empty family Φ_i, $i \in I$ are given by:

$$\Big(\inf_{i\in I} \Phi_i\Big)(A) = \bigcap_{i\in I} \Phi_i(A) \text{ and } \Big(\sup_{i\in I} \Phi_i\Big)(A) = \bigcup_{i\in I} \Phi_i(A) \text{ for all category sets } A.$$

We denote by $\mathbb{C}$ the set of all coherent inference systems:

$$\mathbb{C} \coloneqq \{\Phi \in \mathbb{S} \colon (\forall A \in \mathbb{F})\Phi(A) \text{ is Bernstein coherent}\}. \tag{24}$$

Then it is clear that $\mathbb{C}$ is a complete meet-semilattice, meaning that it is closed under arbitrary non-empty infima:[27]

$$(\forall i \in I)\Phi_i \in \mathbb{C} \Rightarrow \inf_{i\in I} \Phi_i \in \mathbb{C}. \tag{25}$$

The bottom of this structure—the most conservative coherent inference system—is called the *vacuous inference system* Φ_{V}, and it is the coherent inference system given by:

$$\Phi_{\mathrm{V}}(A) = \mathcal{V}^+(A) \text{ for all category sets } A.$$

We shall come back in some detail to this vacuous inference system in Section 9.

The property (25) allows us to do conservative reasoning with coherent inference systems. Suppose, for instance, that for some collection of category sets $\mathcal{F} \subseteq \mathbb{F}$, we have assessments

27. It is not necessarily closed under suprema, however, as the union of Bernstein coherent sets of polynomials need not be Bernstein coherent.

$\mathcal{A}$ in the form of a set of polynomials $\mathcal{A}_A \subseteq \mathcal{V}(A)$, $A \in \mathcal{F}$. Then, if it exists, the most conservative coherent inference system $\Phi_{\mathcal{A}}$ that is compatible with these assessments is given by:

$$\Phi_{\mathcal{A}} = \inf\left\{\Phi \in \mathbb{C}\colon (\forall A \in \mathcal{F})\mathcal{A}_A \subseteq \Phi(A)\right\}.$$

And, of course, it will exist if and only the set of polynomials $\mathcal{A}_A$ is included in some Bernstein coherent set of polynomials $\mathcal{H}_A$ on A, for all $A \in \mathcal{F}$. In that case, it is not difficult to see, given the discussion in Section 5.3, that $\Phi_{\mathcal{A}}(A) = \operatorname{posi}(\mathcal{V}^+(A) \cup \mathcal{A}_A)$ for $A \in \mathcal{F}$ and $\Phi_{\mathcal{A}}(A) = \mathcal{V}^+(A)$ for $A \in \mathbb{F} \setminus \mathcal{F}$.

7. Representation Insensitivity and Specificity under Exchangeability

Let us now investigate what form the inference principles of representation insensitivity (RI2) and specificity (SP2) take for predictive inference under exchangeability, when such inference can be completely characterised by Bernstein coherent sets of polynomials. This will allow us to reformulate these principles as constraints on—or properties of—inference systems.

7.1 Representation Insensitivity

We recall the notations and assumptions in Section 4.4. With the surjective (onto) map $\rho\colon A \to D$ we associate the surjective map $R_\rho\colon \mathbb{R}^A \to \mathbb{R}^D$ by letting:

$$R_\rho(\boldsymbol{\alpha})_z := \sum_{x \in A\colon \rho(x)=z} \alpha_x \quad \text{for all } \boldsymbol{\alpha} \in \mathbb{R}^A \text{ and all } z \in D. \tag{26}$$

This map allows us to give the following elegant characterisation of representation insensitivity.

Theorem 7. *A coherent inference system Φ is representation insensitive if and only if for all category sets A and D such that there is an onto map $\rho\colon A \to D$, for all $p \in \mathcal{V}(D)$ and all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:*

$$(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \Phi(D). \tag{RI3}$$

Running Example. Assume now that the coins in the bag are actually rather thick, implying that there is a non-negligible chance that they do not fall on one of their flat sides, but remain upright. If we denote this new 'up' state by U, then we have a new category set $A := \{H, T, U\}$. If we also consider a new 'flat' state F, meaning either heads or tails, then we can also consider, instead of A, the category set $D := \{F, U\}$ that does not distinguish between heads and tails. The relabelling map ρ with $\rho(H) := \rho(T) := F$ and $\rho(U) := U$ identifies the proper relations between the categories in A and D.

Suppose now that we want to say something about the lower probability of the event $\widehat{UF}$ of observing U on one flip and H or T on the other, immediately after observing the sequence (H, U, H, T) with count vector $\boldsymbol{m} = (2, 1, 1)$—the last count in three from now on refers to the number of Us in the observation sequence. In the A-domain, the gamble $\mathbb{I}_{\widehat{UF}}$ can be expressed by the polynomial $q = \mathrm{Mn}^{\hat{n}}_{\{H,T,U\}}(\mathbb{I}_{\widehat{UF}})$, $\hat{n} \geq 2$ given by:[28]

$$q(\boldsymbol{\theta}) = 2(\theta_H + \theta_T)\theta_U \text{ for all } \boldsymbol{\theta} \in \Sigma_{\{H,T,U\}}.$$

28. As before in similar contexts, it is easy to check that this polynomial remains the same for all $\hat{n} \geq 2$.

So we want to find out whether polynomials of the type $[2(\theta_H+\theta_T)\theta_U-\mu]12\theta_H^2\theta_T\theta_U$ belong to $\Phi(\{H,T,U\})$; see Equation (18).

On the other hand, as we have seen previously, in the D-domain, the gamble $\mathbb{I}_{\widehat{UF}}$ can be expressed by the polynomial p given by $p(\boldsymbol{\vartheta})=2\vartheta_F\vartheta_U$ for all $\boldsymbol{\vartheta}\in\Sigma_{\{F,U\}}$. Observe that $q=p\circ R_\rho$. The count vector $\boldsymbol{m}=(2,1,1)$ in the A-domain corresponds to a count vector $R_\rho(\boldsymbol{m})=(3,1)$ in the D-domain, where the first component refers to the number of Fs and the second to the number of Us. So here, we need to check whether polynomials of the type $[2\vartheta_F\vartheta_U-\mu]4\vartheta_F^3\vartheta_U$ belong to $\Phi(\{F,U\})$.

The nice thing about representation insensitivity is that it makes checking whether polynomials of the type $[2(\theta_H+\theta_T)\theta_U-\mu]12\theta_H^2\theta_T\theta_U$ belong to $\Phi(\{H,T,U\})$ in the A-domain equivalent to checking whether polynomials of the type $[2\vartheta_F\vartheta_U-\mu]4\vartheta_F^3\vartheta_U$ belong to $\Phi(\{F,U\})$ in the D-domain. $\Diamond$

Very interestingly, representation insensitivity is preserved under taking arbitrary non-empty infima of coherent inference systems, which allows us to look for the most conservative representation insensitive coherent inference system that is compatible with an assessment $\mathcal{A}$ on $\mathcal{F}$, in a way that is a straightforward extension of the discussion near the end of Section 6.

Theorem 8. *Consider any non-empty family Φ_i, $i\in I$ of representation insensitive coherent inference systems. Then their infimum $\inf_{i\in I}\Phi_i$ is a representation insensitive coherent inference system as well.*

7.2 Specificity

Next, we turn to specificity, and recall the notations and assumptions in Section 4.5. Let us define the surjective *restriction map* $r_B\colon\mathbb{R}^A\to\mathbb{R}^B$ by:

$$r_B(\boldsymbol{\alpha})_z:=\alpha_z \text{ for all } \boldsymbol{\alpha}\in\mathbb{R}^A \text{ and all } z\in B, \tag{27}$$

so in particular, $r_B(\boldsymbol{m})$ is the count vector on B obtained by restricting to B the (indices of the) components of the count vector $\boldsymbol{m}$ on A. We also define the one-to-one *injection map* $i_A\colon\mathbb{R}^B\to\mathbb{R}^A$ by:

$$i_A(\boldsymbol{\alpha})_x:=\begin{cases}\alpha_x & \text{if } x\in B\\ 0 & \text{otherwise}\end{cases}\quad \text{for all } \boldsymbol{\alpha}\in\mathbb{R}^B \text{ and all } x\in A. \tag{28}$$

This map can be used to define the following one-to-one maps $\mathrm{I}^r_{B,A}\colon\mathcal{V}(B)\to\mathcal{V}(A)$, for any $r\in\mathbb{N}_0$, as follows:

$$\mathrm{I}^r_{B,A}(p):=\sum_{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}} b_p^{\deg(p)+r}(\boldsymbol{n})\mathrm{B}_{A,i_A(\boldsymbol{n})} \text{ for all polynomials } p \text{ in } \mathcal{V}(B). \tag{29}$$

They derive their meaning from the following observation. A polynomial p on Σ_B can be equivalently represented in any Bernstein basis on Σ_B of degree $\deg(p)+r$. But when we interpret these different representations as polynomials on Σ_A, they are no longer equivalent, and lead to different polynomials $\mathrm{I}^r_{B,A}(p)$, $r\in\mathbb{N}_0$. The following propositions clarify what exactly the effect of the operator $\mathrm{I}^r_{B,A}$ is.

Proposition 9. *For any polynomial p on Σ_B and any $r \in \mathbb{N}_0$: $\mathrm{I}^r_{B,A}(p) \circ i_A = p$.*

We introduce the following notation, for any $\boldsymbol{\theta} \in \Sigma_A$ such that $\theta_B > 0$: $\boldsymbol{\theta}|^+_B \coloneqq r_B(\boldsymbol{\theta})/\theta_B$. Observe that $\boldsymbol{\theta}|^+_B \in \Sigma_B$ whenever $\theta_B > 0$.

Proposition 10. *Consider any polynomial p on Σ_B, any $r \in \mathbb{N}_0$ and any $\boldsymbol{\theta} \in \Sigma_A$. When $\deg(p) + r = 0$ then $p = c \in \mathbb{R}$, and $\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) = c$. Otherwise, when $\deg(p) + r > 0$:*

$$\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) = \begin{cases} \theta_B^{\deg(p)+r} p(\boldsymbol{\theta}|^+_B) & \text{if } \theta_B > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The maps $\mathrm{I}^r_{B,A}$ allow us to give the following elegant characterisation of specificity:

Theorem 11. *A coherent inference system Φ is specific if and only if for all category sets A and B such that $B \subseteq A$, for all $p \in \mathcal{V}(B)$, all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and all $r \in \mathbb{N}_0$:*

$$\mathrm{I}^r_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \Phi(B). \tag{SP3}$$

Running Example. Suppose, as before, that we have made the observation (H, U, H, T), with count vector $\boldsymbol{m} = (2, 1, 1)$. We are interested in the posterior lower probability of the event $\widehat{HT}$, where somebody has told us that neither of the two subsequent coin flips—after the first four—resulted in U. In a specific inference system, we are allowed to consider this predictive inference problem in the reduced category space $B = \{H, T\}$, rather than in the category space $A = \{H, T, U\}$. But then, in the B-space, we have to use the reduced count vector $r_B(\boldsymbol{m}) = (2, 1)$, obtained by leaving out the number of observed Us. The polynomials we are lead to consider here, are therefore of the type $[2\vartheta_H\vartheta_T - \mu]3\vartheta_H^2\vartheta_T$, for which we want to know whether they belong to $\Phi(\{F, U\})$.

In the A-space, the polynomial $p(\boldsymbol{\vartheta}) = 2\vartheta_H\vartheta_T - \mu$, whose degree is $\deg(p) = 2$, is transformed into the polynomials

$$\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) = \left[2\frac{\theta_H}{\theta_H + \theta_T}\frac{\theta_T}{\theta_H + \theta_T} - \mu\right](\theta_H + \theta_T)^{2+r} = [2\theta_H\theta_T - \mu(\theta_H + \theta_T)^2](\theta_H + \theta_T)^r$$

for $r \in \mathbb{N}_0$. It follows from the argumentation in the proof of Theorem 11 that the original problem requires us to check whether polynomials of the type

$$[2\theta_H\theta_T - \mu(\theta_H + \theta_T)^2](\theta_H + \theta_T)^r 12\theta_H^2\theta_T\theta_U$$

are in $\Phi(\{H, T, U\})$. Specificity allows us to look at the problem in the B-space, which is easier. $\lozenge$

Observe the close formal similarity between the conditions (RI3) and (SP3). It should therefore not surprise us that specificity, too, is preserved under taking arbitrary non-empty infima of inference systems.

Theorem 12. *Consider any non-empty family Φ_i, $i \in I$ of specific coherent inference systems. Then their infimum $\inf_{i \in I} \Phi_i$ is a specific coherent inference system as well.*

Let us denote by $\mathbb{C}_{\mathrm{rs}}$ the set of all coherent inference systems that are both representation insensitive and specific. It follows from Theorems 8 and 12 that $\mathbb{C}_{\mathrm{rs}}$, *like $\mathbb{C}$, is closed under arbitrary non-empty infima, so we can perform conservative reasoning, in very much the same way as we discussed near the end of Section 6.*

8. Immediate Prediction

If we have an inference system Φ, we can look at the special case of *immediate prediction*, where for a given category set A, after observing a sample of $\check{n} \geq 0$ variables with count vector $\check{\boldsymbol{m}} \in \mathcal{N}_A^{\check{n}}$, we want to express beliefs about the value that the next observation $X_{\check{n}+1}$ will assume in A. So this is the specific case of predictive inference with $\hat{n} = 1$, and Condition (18) can now be simplified somewhat, as for all gambles f on A and all $\check{\boldsymbol{m}} \in \mathcal{N}_A^{\check{n}}$:

$$f \in \mathcal{D}_A^1 \Leftrightarrow \mathrm{S}_A(f) \in \Phi(A) \text{ and } f \in \mathcal{D}_A^1 \rfloor \check{\boldsymbol{m}} \Leftrightarrow \mathrm{B}_{A,\check{\boldsymbol{m}}}\, \mathrm{S}_A(f) \in \Phi(A),$$

where we let the so-called *sampling expectation* $\mathrm{S}_A(f)$ be the linear polynomial on Σ_A given by $\mathrm{S}_A(f|\boldsymbol{\theta}) \coloneqq \sum_{x \in A} f(x)\theta_x$ for all $\boldsymbol{\theta} \in \Sigma_A$.

The reason for this is that $\mathcal{N}_A^1 = \{\boldsymbol{e}^x : x \in A\}$ where $\boldsymbol{e}^x$ is the count vector corresponding to a single observation of category x, or in other words, $e_z^x = \delta_{xz}$ for all $z \in A$ [Kronecker delta]. Hence, for all $x \in A$ and any $\boldsymbol{\theta} \in \Sigma_A$:

$$\mathrm{Hy}_A^1(f|\boldsymbol{e}^x) = \binom{1}{\boldsymbol{e}^x} \sum_{z \in [\boldsymbol{e}^x]} f(z) = f(x) \text{ and } \mathrm{B}_{A,\boldsymbol{e}^x}(\boldsymbol{\theta}) = \binom{1}{\boldsymbol{e}^x} \prod_{z \in A} \theta_z^{e_z^x} = \theta_x,$$

leading to:

$$\mathrm{Mn}_A^1(f|\boldsymbol{\theta}) = \sum_{\boldsymbol{e}^x \in \mathcal{N}_A^1} \mathrm{Hy}_A^1(f|\boldsymbol{e}^x)\mathrm{B}_{A,\boldsymbol{e}^x}(\boldsymbol{\theta}) = \sum_{x \in A} f(x)\theta_x = \mathrm{S}_A(f|\boldsymbol{\theta}). \tag{30}$$

It is a matter of straightforward verification that, due to the Bernstein coherence of $\mathcal{H}_A$, the so-called *immediate prediction model* $\mathcal{D}_A^1 \rfloor \check{\boldsymbol{m}}$ is a coherent set of desirable gambles on A, for every count vector $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. It induces the following predictive lower previsions:

$$\begin{aligned}\underline{P}_A^1(f|\check{\boldsymbol{m}}) &= \sup\left\{\alpha \in \mathbb{R} \colon f - \alpha \in \mathcal{D}_A^1 \rfloor \check{\boldsymbol{m}}\right\} \\ &= \sup\left\{\alpha \in \mathbb{R} \colon [\mathrm{S}_A(f) - \alpha]\mathrm{B}_{A,\check{\boldsymbol{m}}} \in \Phi(A)\right\}.\end{aligned} \tag{31}$$

Immediate prediction in the context of exchangeable imprecise probability models has been studied in some detail by De Cooman et al. (2009a). Lower previsions, rather than sets of desirable gambles, were the model of choice in that paper, and because of that, the authors encountered problems with conditioning on sets of (lower) probability zero. In fact, it is these problems that provided the motivation for dealing with the much more general problem of (not necessarily immediate) predictive inference using sets of desirable gambles in the present paper. In this section, we want to illustrate how many of the results proved there can be made stronger (and with easier proofs, as is borne out in Appendix E.3) in the present context.

The requirement (RI2) for representation insensitivity reduces to the following simpler requirement on immediate prediction models: for all category sets A and D such that there is an onto map $\rho \colon A \to D$, for all gambles f on D and all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$:

$$f \circ \rho \in \mathcal{D}_A^1 \rfloor \boldsymbol{m} \Leftrightarrow f \in \mathcal{D}_D^1 \rfloor R_\rho(\boldsymbol{m}). \tag{RI4}$$

Similarly, the requirement (SP2) for specificity reduces to the following simpler requirement on immediate prediction models: for all category sets A and B such that $B \subseteq A$, for all gambles f on B and all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$:

$$f\mathbb{I}_B \in \mathcal{D}_A^1 \rfloor \boldsymbol{m} \Leftrightarrow f \in \mathcal{D}_B^1 \rfloor r_B(\boldsymbol{m}). \tag{SP4}$$

Let us now show that there is a simple characterisation of the immediate prediction models that satisfy representation insensitivity. To get there, observe that we can consider any gamble g on a category set A as a (surjective) pooling map from A to the finite subset $g(A)$ of $\mathbb{R}$—also a category set. The corresponding $R_g\colon \mathbb{R}^A \to \mathbb{R}^{g(A)}$ is given by:

$$R_g(\boldsymbol{\alpha})_r = \sum_{x\in A\colon g(x)=r} \alpha_x \text{ for all } r \in g(A).$$

This simple idea allows for an intriguing reformulation of the representation insensitivity requirement on immediate prediction models:

Proposition 13. *The immediate prediction models associated with a coherent inference system are representation insensitive if and only if for all category sets A, all gambles g on A and all count vectors $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:*

$$g \in \mathcal{D}^1_A \rfloor \boldsymbol{m} \Leftrightarrow \mathrm{id}_{g(A)} \in \mathcal{D}^1_{g(A)} \rfloor R_g(\boldsymbol{m}). \tag{RI5}$$

Here, for any non-empty set B, we denote by id_B the identity map on B, defined by $\mathrm{id}_B(z) \coloneqq z$ for all $z \in B$.

Proposition 13 tells us that whether a gamble is desirable depends only on the values it assumes—and not on where they are assumed—and on the number of times each of these values has been observed in the past—or rather would have been if we had been observing the $g(X_k)$ rather than the X_k.

Let us now focus on what happens for events. Consider any event $B \subseteq A$ that is *non-trivial*—meaning that B is neither empty nor equal to A. Then for any real μ the gamble $\mathbb{I}_B - \mu$ assumes two values, $1-\mu$ and $-\mu$, so we see after applying Proposition 13 that for all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathbb{I}_B - \mu \in \mathcal{D}^1_A \rfloor \boldsymbol{m} \Leftrightarrow \mathrm{id}_{\{1-\mu,-\mu\}} \in \mathcal{D}^1_{\{1-\mu,-\mu\}} \rfloor (m_B, m_{A\setminus B}),$$

and therefore

$$\underline{P}^1_A(B|\boldsymbol{m}) = \sup\left\{\mu \in \mathbb{R}\colon \mathbb{I}_B - \mu \in \mathcal{D}^1_A \rfloor \boldsymbol{m}\right\} \tag{32}$$

$$= \sup\left\{\mu \in \mathbb{R}\colon \mathrm{id}_{\{1-\mu,-\mu\}} \in \mathcal{D}^1_{\{1-\mu,-\mu\}} \rfloor (m_B, m_{A\setminus B})\right\} =\colon \varphi(m_A, m_B), \tag{33}$$

meaning that, by representation insensitivity, the predictive lower probability for a *non-trivial* event B depends only on the number of times m_B that it has been observed in the past experiments, and the total number of observations m_A. The same thing holds for its predictive upper probability $1 - \varphi(m_A, m_A - m_B)$. For precise predictive probabilities, a similar property is known as *Johnson's sufficientness postulate* (Johnson, 1924; Zabell, 1982).

So for any representation insensitive coherent inference system, we see that we can define a so-called *lower probability function* $\varphi\colon \left\{(n,k) \in \mathbb{N}_0^2\colon k \leq n\right\} \to [0,1]$ through Equation (33), which completely characterises the 'one-step-ahead' predictive lower and upper probabilities[29] for all non-trivial events and all count vectors. We shall now use the representation insensitivity and specificity requirements to try and say more about this lower probability function. The following theorem strengthens, simplifies, and extends similar results by De Cooman et al. (2009a).

29. ... but not necessarily the predictive lower and upper previsions ...

Theorem 14. *Consider a representation insensitive coherent inference system Φ. Then the associated lower probability function φ has the following properties:*

L1. *φ is bounded: $0 \leq \varphi(n,k) \leq 1$ for all $n,k \in \mathbb{N}_0$ such that $k \leq n$.*

L2. *φ is super-additive in its second argument: $\varphi(n,k+\ell) \geq \varphi(n,k)+\varphi(n,\ell)$ for all $n,k,\ell \in \mathbb{N}_0$ such that $k+\ell \leq n$.*

L3. *$\varphi(n,0) = 0$ for all $n \in \mathbb{N}_0$.*

L4. *$\varphi(n,k) \geq k\varphi(n,1)$ and $n\varphi(n,1) \leq 1$ for all $n,k \in \mathbb{N}_0$ such that $1 \leq k \leq n$.*

L5. *φ is non-decreasing in its second argument: $\varphi(n,k) \leq \varphi(n,\ell)$ for all $n,k,\ell \in \mathbb{N}_0$ such that $k \leq \ell \leq n$.*

L6. *$\varphi(n,k) \geq \varphi(n+1,k) + \varphi(n,k)[\varphi(n+1,k+1) - \varphi(n+1,k)]$ for all $n,k \in \mathbb{N}_0$ such that $k \leq n$.*

L7. *φ is non-increasing in its first argument: $\varphi(n+1,k) \leq \varphi(n,k)$ for all $n,k \in \mathbb{N}_0$ such that $k \leq n$.*

L8. *Suppose that $\varphi(n,1) > 0$ for all $n \in \mathbb{N}$, and let $s_n := \frac{1}{\varphi(n,1)} - n$, or equivalently, $\varphi(n,1) = \frac{1}{n+s_n}$. Then $s_n \geq 0$ and $s_{n+1} \geq s_n$.*

If Φ is moreover specific, then φ has the following properties:

L9. *Consider any real $\alpha \in (0,1)$ and suppose that $\varphi(1,1) \geq \alpha$, then $\varphi(n,n) \geq \frac{n\alpha}{1-\alpha+n\alpha}$ for all $n \in \mathbb{N}_0$. As a consequence, consider any $s > 0$ and suppose that $\varphi(1,1) \geq \frac{1}{1+s}$, then $\varphi(n,n) \geq \frac{n}{n+s}$ for all $n \in \mathbb{N}_0$.*

We know from Theorem 4 that representation insensitive coherent inference systems are near-ignorant, meaning that they are vacuous and therefore completely indecisive about any single observation when no prior observations have been made. This is also borne out by Theorem 14.L3. Let us define the *imprecision function* by

$$\iota(n,k) := 1 - \varphi(n,n-k) - \varphi(n,k) \text{ for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n. \tag{34}$$

It is clear that $\overline{P}^1_A(B|\boldsymbol{m}) - \underline{P}^1_A(B|\boldsymbol{m}) = \iota(m_A, m_B)$ is the width of the probability interval for an event B that has been observed before m_B out of m_A times. For a representation insensitive coherent inference system whose imprecision function $\iota(n,k)$ satisfies the following property:

$$\left.\begin{aligned} \iota(n+1,k) &\leq \iota(n,k) \\ \iota(n+1,k+1) &\leq \iota(n,k) \end{aligned}\right\} \text{ for all } 0 \leq k \leq n, \tag{35}$$

the imprecision does not increase as the total number of observations increases. This suggests that such representation insensitive coherent inference systems will display some of the desirable behaviour mentioned in the Introduction: they are conservative when little has been learned, and they never become less precise as more observations come in. In the following sections, we intend—amongst other things—to take a closer look at whether this behaviour is present in a number of such systems.

Immediate prediction is very important for predictive inference with precise probabilities, as the Law of Total Probability guarantees that it is completely determined by its immediate predictions. Perhaps surprisingly, this is not the case for predictive inference with imprecise probabilities: Appendix D provides a counterexample. This also points to some of the limitations in scope of the earlier work by De Cooman et al. (2009a). For this reason, we now leave immediate prediction models for what they are, and in the rest of this paper concentrate on the more general notion of an inference system.

9. The Vacuous Inference System

In this and the following sections, we provide explicit and interesting examples of representation insensitive, and of specific coherent inference systems. We begin with the simplest one: the vacuous inference system Φ_{V}, which we introduced in Section 6 as the smallest, or most conservative, coherent inference system. It associates with any category set A the smallest Bernstein coherent set $\Phi_{\mathrm{V}}(A)=\mathcal{H}_{\mathrm{V},A}:=\mathcal{V}^+(A)$ containing all the Bernstein positive polynomials—the ones that are guaranteed to be there anyway, by Bernstein coherence alone. We deduce from Proposition 30 in Appendix B that:

$$\mathcal{H}_{\mathrm{V},A}\rfloor\check{\boldsymbol{m}}=\mathcal{H}_{\mathrm{V},A}=\mathcal{V}^+(A) \text{ for all } \check{\boldsymbol{m}}\in\mathcal{N}_A\cup\{\mathbf{0}\},$$

and from Proposition 28 in Appendix B that:

$$\begin{aligned}\underline{H}_{\mathrm{V},A}(p|\check{\boldsymbol{m}})=\underline{H}_{\mathrm{V},A}(p)&=\sup\left\{\alpha\in\mathbb{R}\colon p-\alpha\in\mathcal{V}^+(A)\right\}\\&=\min p=\min_{\boldsymbol{\theta}\in\Sigma_A}p(\boldsymbol{\theta}) \text{ for all } p\in\mathcal{V}(A).\end{aligned}$$

The predictive models for this inference system are now straightforward to find, as they follow directly from Equations (21) and (23). For any $\hat{n}\in\mathbb{N}$ and any $\check{\boldsymbol{m}}\in\mathcal{N}_A\cup\{\mathbf{0}\}$, we deduce that:

$$\mathcal{D}^{\hat{n}}_{\mathrm{V},A}=\mathcal{D}^{\hat{n}}_{\mathrm{V},A}\rfloor\check{\boldsymbol{m}}=\left\{f\in\mathcal{L}(A^{\hat{n}})\colon\ \mathrm{Mn}^{\hat{n}}_A(f)\in\mathcal{V}^+(A)\right\},\tag{36}$$

and

$$\underline{P}^{\hat{n}}_{\mathrm{V},A}(f)=\underline{P}^{\hat{n}}_{\mathrm{V},A}(f|\check{\boldsymbol{m}})=\min_{\boldsymbol{\theta}\in\Sigma_A}\mathrm{Mn}^{\hat{n}}_A(f|\boldsymbol{\theta}) \text{ for all } f\in\mathcal{L}(A^{\hat{n}}).\tag{37}$$

In particular:

$$\mathcal{D}^1_{\mathrm{V},A}=\mathcal{D}^1_{\mathrm{V},A}\rfloor\check{\boldsymbol{m}}=\mathcal{L}_{>0}(A),\tag{38}$$

$$\underline{P}^1_{\mathrm{V},A}(f)=\underline{P}^1_{\mathrm{V},A}(f|\check{\boldsymbol{m}})=\min f \text{ for all } f\in\mathcal{L}(A),\tag{39}$$

and

$$\varphi_{\mathrm{V}}(n,k)=0 \text{ for all } n,k\in\mathbb{N}_0 \text{ such that } k\leq n.\tag{40}$$

These are the most conservative exchangeable predictive models there are, and they arise from making no other assessments than exchangeability alone. As we gather from Equations (36)–(40), they are not very interesting, because they involve no non-trivial commitments, and they do not allow for learning from observations. This is also borne out by the corresponding imprecision function, which is given by:

$$\iota_{\mathrm{V}}(n,k)=1 \text{ for all } n,k\in\mathbb{N}_0 \text{ such that } k\leq n.$$

Running Example. We have seen before that $\mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = 2\theta_H\theta_T$ for all $\hat{n} \geq 2$, and therefore

$$\underline{P}^{\hat{n}}_{\mathrm{V},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \underline{P}^{\hat{n}}_{\mathrm{V},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \min_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} \mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} 2\theta_H\theta_T = 0$$

and

$$\overline{P}^{\hat{n}}_{\mathrm{V},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \overline{P}^{\hat{n}}_{\mathrm{V},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \max_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} \mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} 2\theta_H\theta_T = \frac{1}{2}.$$

This shows that the vacuous inference model does not produce completely vacuous inferences: it allows us to find out the consequences of making no other assessments than exchangeability. But it does not allow us to change our lower and upper probabilities and previsions when new observations come in. ◊

Even though it makes no non-trivial inferences, the vacuous inference system satisfies representation insensitivity, but it is not specific.

Theorem 15. *The vacuous inference system Φ_{V} is coherent and representation insensitive.*

Let us show by means of a counterexample that Φ_{V} is not specific,

Running Example. Let us go back to inferences about the category space $A = \{H, T, U\}$ and the reduced category space $B = \{H, T\}$. Consider the polynomial $p(\boldsymbol{\vartheta}) = \vartheta_H^2 - \vartheta_H\vartheta_T + \vartheta_T^2$ on $\Sigma_{\{H,T\}}$. This polynomial is Bernstein positive—so $p \in \mathcal{V}^+(\{H, T\})$—because

$$p(\boldsymbol{\vartheta}) = (\vartheta_H^2 - \vartheta_H\vartheta_T + \vartheta_T^2)(\vartheta_H + \vartheta_T) = \vartheta_H^3 + \vartheta_T^3$$

has an expansion in the Bernstein basis of degree 3 that is positive. But let us now consider the corresponding polynomial on $\Sigma_{\{H,T,U\}}$:

$$q(\boldsymbol{\theta}) := \mathrm{I}^0_{B,A}(p|\boldsymbol{\theta}) = \theta_H^2 - \theta_H\theta_T + \theta_T^2. \tag{41}$$

This polynomial is not Bernstein positive: it is easy to see that for every $n \in \mathbb{N}_0$,

$$q(\boldsymbol{\theta}) = (\theta_H^2 - \theta_H\theta_T + \theta_T^2)(\theta_H + \theta_T + \theta_U)^n$$

will always have a term $-\theta_H\theta_T\theta_U^n$. So $q = \mathrm{I}^0_{B,A}(p) \notin \mathcal{V}^+(\{H, T, U\})$, and we infer from Theorem 11 that Φ_{V} cannot be specific. ◊

In the following sections, we shall prove that there are an infinity of more committal, specific and representation insensitive coherent inference systems. We begin by introducing a slightly modified version of the vacuous inference system that is coherent, representation insensitive *and specific.*

10. The Nearly Vacuous Inference System

Let us introduce the *nearly vacuous* inference system Φ_{NV}—the reason for its name will become clear presently—by:

$$\Phi_{\mathrm{NV}}(A) \coloneqq \mathcal{H}_{\mathrm{NV},A} \coloneqq \mathcal{V}^{++}(A) \coloneqq \{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A))p(\boldsymbol{\theta}) > 0\}$$
for all category sets A.

Since $\mathcal{V}^{++}(A)$ consists of all the polynomials that are positive on $\mathrm{int}(\Sigma_A)$, we deduce from Proposition 28 in Appendix B that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$: $\mathcal{H}_{\mathrm{NV},A}\rfloor\check{\boldsymbol{m}} = \mathcal{H}_{\mathrm{NV},A} = \mathcal{V}^{++}(A)$ and that:

$$\underline{H}_{\mathrm{NV},A}(p|\check{\boldsymbol{m}}) = \underline{H}_{\mathrm{NV},A}(p) = \inf_{\boldsymbol{\theta}\in\mathrm{int}(\Sigma_A)} p(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Sigma_A} p(\boldsymbol{\theta}) \text{ for all } p \in \mathcal{V}(A).$$

Since we know from Proposition 28 in Appendix B, and the counterexample following it, that generally speaking $\mathcal{V}^{+}(A) \subset \mathcal{V}^{++}(A)$, we see that this inference system is less conservative than the vacuous one. As was the case for the vacuous inference system, the predictive models for this nearly vacuous inference system are straightforward to find, as they follow directly from Equations (21) and (23). For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$, we deduce that:

$$\mathcal{D}^{\hat{n}}_{\mathrm{NV},A} = \mathcal{D}^{\hat{n}}_{\mathrm{NV},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A^{\hat{n}}) \colon\ \mathrm{Mn}^{\hat{n}}_A(f) \in \mathcal{V}^{++}(A)\right\},$$

and

$$\underline{P}^{\hat{n}}_{\mathrm{NV},A}(f) = \underline{P}^{\hat{n}}_{\mathrm{NV},A}(f|\check{\boldsymbol{m}}) = \min_{\boldsymbol{\theta}\in\Sigma_A} \mathrm{Mn}^{\hat{n}}_A(f|\boldsymbol{\theta}) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}).$$

In particular:

$$\mathcal{D}^{1}_{\mathrm{NV},A} = \mathcal{D}^{1}_{\mathrm{NV},A}\rfloor\check{\boldsymbol{m}} = \mathcal{L}_{>0}(A),$$
$$\underline{P}^{1}_{\mathrm{NV},A}(f) = \underline{P}^{1}_{\mathrm{NV},A}(f|\check{\boldsymbol{m}}) = \min f \text{ for all } f \in \mathcal{L}(A).$$

We see that the immediate prediction models, and the predictive lower previsions, for this inference system are exactly the same as the ones for the vacuous inference systems.[30] They too do not allow for learning from observations.

Interestingly, and in contrast with the vacuous inference system, the nearly vacuous inference system is specific, which already tells us that $\mathbb{C}_{\mathrm{rs}} \neq \emptyset$.

Theorem 16. *The nearly vacuous inference system Φ_{NV} is coherent, representation insensitive and specific: $\Phi_{\mathrm{NV}} \in \mathbb{C}_{\mathrm{rs}}$.*

11. The Skeptically Cautious Inference System

We now construct a rather simple inference system that is quite intuitive and slightly more informative than the vacuous and nearly vacuous ones. Suppose that our subject uses the following system for making inferences based on a sequence of $\check{n} > 0$ observations with count vector $\check{\boldsymbol{m}}$, in a category set A. He is 'skeptical' in that he believes that in the future, he will only observe categories that he has seen previously, so only categories in the set:

$$A[\check{\boldsymbol{m}}] \coloneqq \{x \in A \colon \check{m}_x > 0\}\,. \tag{42}$$

But he is also 'cautious', because his beliefs about which of these already observed categories will be observed in the future, are 'nearly' vacuous. To explain this, assume first that in

30. This is a first example that shows that the immediate prediction models do not completely determine the inference system. We shall come across another example in Appendix D.

particular, for $\hat{n}$ future observations, he has vacuous beliefs about which count vector he will observe in the set

$$\left\{\hat{\boldsymbol{m}} \in \mathcal{N}_A^{\hat{n}} \colon (\forall y \in A \setminus A[\check{\boldsymbol{m}}])\hat{m}_y = 0\right\} = \mathcal{N}_{A[\check{\boldsymbol{m}}]}^{\hat{n}}$$

of those future count vectors $\hat{\boldsymbol{m}}$ that he holds possible after observing the count vector $\check{\boldsymbol{m}}$, namely those count vectors with no observation outside $A[\check{\boldsymbol{m}}]$.[31] By Lemma 47 in Appendix B, this would lead us to associate the following set of polynomials with any count vector $\boldsymbol{m} \in \mathcal{N}_A$:

$$\begin{aligned}\mathcal{V}^+_{[\boldsymbol{m}]}(A) &\coloneqq \left\{p \in \mathcal{V}(A) \colon (\exists n \geq \deg(p))\, b^n_{p\,|\mathcal{N}^n_{A[\boldsymbol{m}]}} > 0\right\} \\ &= \left\{p \in \mathcal{V}(A) \colon p_{|\Sigma_{A[\boldsymbol{m}]}} \in \mathcal{V}^+(A[\boldsymbol{m}])\right\}.\end{aligned}$$

But, because we already know that the vacuous models $\mathcal{V}^+(A)$ do not lead to specific systems, whereas the nearly vacuous models $\mathcal{V}^{++}(A)$ do, we will modify this slightly, and rather associate the following set of polynomials with any count vector $\boldsymbol{m} \in \mathcal{N}_A$:

$$\mathcal{V}^{++}_{[\boldsymbol{m}]}(A) \coloneqq \left\{p \in \mathcal{V}(A) \colon p_{|\Sigma_{A[\boldsymbol{m}]}} \in \mathcal{V}^{++}(A[\boldsymbol{m}])\right\}.$$

The polynomials in $\mathcal{V}^{++}_{[\boldsymbol{m}]}(A)$ are 'desirable in representation'[32] after observing a sample with count vector $\boldsymbol{m}$, so we infer from Equation (19) that the subject considers as 'desirable in representation' all polynomials in:

$$\mathcal{V}^{++}_{[\boldsymbol{m}]}(A)\mathrm{B}_{A,\boldsymbol{m}} = \left\{p\mathrm{B}_{A,\boldsymbol{m}} \colon p \in \mathcal{V}^{++}_{[\boldsymbol{m}]}(A)\right\}.$$

We are thus led to consider the following assessment:

$$\mathcal{A}_{\mathrm{SC},A} \coloneqq \bigcup_{\boldsymbol{m} \in \mathcal{N}_A} \mathcal{V}^{++}_{[\boldsymbol{m}]}(A)\mathrm{B}_{A,\boldsymbol{m}},$$

and the set of all its positive linear combinations:

$$\mathcal{H}_{\mathrm{SC},A} \coloneqq \operatorname{posi}(\mathcal{A}_{\mathrm{SC},A}) = \left\{\sum_{k=1}^{\ell} p_k \mathrm{B}_{A,\boldsymbol{m}_k} \colon \ell \in \mathbb{N}, n_k \in \mathbb{N}, \boldsymbol{m}_k \in \mathcal{N}_A^{n_k}, p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}_k]}(A)\right\}. \quad (43)$$

The following proposition guarantees that the sets $\mathcal{H}_{\mathrm{SC},A}$ are the appropriate most conservative models that summarise the exchangeable inferences for our skeptically cautious subject.

Proposition 17. $\mathcal{H}_{\mathrm{SC},A}$ *is the smallest Bernstein coherent set of polynomials on* Σ_A *that includes* $\mathcal{A}_{\mathrm{SC},A}$.

31. The last equality in the equation above is actually a device that allows us to identify the count vectors on A whose components outside $A[\check{\boldsymbol{m}}]$ are zero, with count vectors on $A[\check{\boldsymbol{m}}]$. We shall be using it repeatedly, without explicit further mention, in the rest of this paper.
32. As stated before, polynomials have no direct behavioural but only an indirect representational meaning, as conveniently condensed representations for desirable gambles on observation sequences. Hence our caution here in using the term 'desirable in representation'.

This also shows that the inference system Φ_{SC}, defined by $\Phi_{\mathrm{SC}}(A) \coloneqq \mathcal{H}_{\mathrm{SC},A}$ for all category sets A, is coherent. We shall call it the *skeptically cautious* inference system.

We now want to find out how updating works in this system. To this end, we introduce a slight generalisation of the set defined in Equation (43). Consider any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$, and let

$$\mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}} \coloneqq \left\{ \sum_{k=1}^{\ell} p_k \mathrm{B}_{A,\boldsymbol{m}_k} : \ell \in \mathbb{N}, n_k \in \mathbb{N}_0, m_A + n_k > 0, \boldsymbol{m}_k \in \mathcal{N}_A^{n_k}, p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}+\boldsymbol{m}_k]}(A) \right\}, \tag{44}$$

so we see that, in particular, $\mathcal{H}_{\mathrm{SC},A} = \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$ for $\boldsymbol{m} = \mathbf{0}$.

The sets $\mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$ have the following interesting characterisation:

Proposition 18. *For all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:*

$$\mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}} = \left\{ p \in \mathcal{V}(A) \setminus \{0\} \colon (\forall K \in \min S_{A,\boldsymbol{m}}(p)) p_{|\Sigma_K} \in \mathcal{V}^{++}(K) \right\}, \tag{45}$$

where

$$S_{A,\boldsymbol{m}}(p) \coloneqq \left\{ \emptyset \neq K \subseteq A \colon A[\boldsymbol{m}] \subseteq K \text{ and } p_{|\Sigma_K} \neq 0 \right\}. \tag{46}$$

By $\min S_{A,\boldsymbol{m}}(p)$ we mean the set of all minimal, or non-dominating, elements of $S_{A,\boldsymbol{m}}(p)$, so $\min S_{A,\boldsymbol{m}}(p) \coloneqq \{C \in S_{A,\boldsymbol{m}}(p) \colon (\forall K \in S_{A,\boldsymbol{m}}(p))(K \subseteq C \Rightarrow K = C)\}$. We formally extend Equation (42) to include the case $\boldsymbol{m} = \mathbf{0}$, so $A[\mathbf{0}] = \emptyset$ and $S_{A,\mathbf{0}}(p) = \left\{ \emptyset \neq K \subseteq A \colon p_{|\Sigma_K} \neq 0 \right\}$.

Proposition 19. *For all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$: $\mathcal{H}_{\mathrm{SC},A} \rfloor \boldsymbol{m} = \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$.*

By combining this result with Equation (21), we can derive—admittedly rather involved—expressions for the predictive sets of desirable gambles for the skeptically cautious inference system. For all $\hat{n} \in \mathbb{N}$ and $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^{\hat{n}}_{\mathrm{SC},A} \rfloor \check{\boldsymbol{m}} = \left\{ f \in \mathcal{L}(A^{\hat{n}}) \colon \operatorname{Mn}^{\hat{n}}_A(f) \in \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}} \right\}. \tag{47}$$

For immediate prediction, these expressions simplify significantly. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\mathcal{D}^1_{\mathrm{SC},A} = \mathcal{L}_{>0}(A) \text{ and } \mathcal{D}^1_{\mathrm{SC},A} \rfloor \check{\boldsymbol{m}} = \left\{ f \in \mathcal{L}(A) \colon f_{|A[\check{\boldsymbol{m}}]} > 0 \right\} \cup \mathcal{L}_{>0}(A). \tag{48}$$

The lower previsions that are derived from $\mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}$ are more tractable. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{H}_{\mathrm{SC},A}(p) = \min_{x \in A} p(\boldsymbol{\theta}^\circ_x) \text{ and } \underline{H}_{\mathrm{SC},A}(p|\check{\boldsymbol{m}}) = \min_{\boldsymbol{\theta} \in \Sigma_{A[\check{\boldsymbol{m}}]}} p(\boldsymbol{\theta}) \text{ for all } p \in \mathcal{V}(A), \tag{49}$$

where, for any $x \in A$, $\boldsymbol{\theta}^\circ_x$ is the degenerate probability mass function on A that assigns all probability mass to x.

The predictive lower previsions for the skeptically cautious inference system are now easily obtained by combining Equations (49) and (23). For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^{\hat{n}}_{\mathrm{SC},A}(f|\check{\boldsymbol{m}}) = \min_{\boldsymbol{\theta} \in \Sigma_{A[\check{\boldsymbol{m}}]}} \operatorname{Mn}^{\hat{n}}_A(f|\boldsymbol{\theta}) \text{ for all } f \in \mathcal{L}(A^n) \tag{50}$$

and

$$\underline{P}^{\hat{n}}_{\mathrm{SC},A}(f) = \min_{x \in A} f(x, x, \dots, x) \text{ for all } f \in \mathcal{L}(A^n). \tag{51}$$

In particular:

$$\underline{P}^{1}_{\mathrm{SC},A}(f) = \min f \text{ and } \underline{P}^{1}_{\mathrm{SC},A}(f|\check{\boldsymbol{m}}) = \min_{x \in A[\check{\boldsymbol{m}}]} f(x) \text{ for all } f \in \mathcal{L}(A). \tag{52}$$

The lower probability function is given by:

$$\varphi_{\mathrm{SC}}(n,k) = \begin{cases} 1 & \text{if } k = n > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n,$$

and the corresponding imprecision function by:

$$\iota_{\mathrm{SC}}(n,k) = \begin{cases} 1 & \text{if } n = 0 \text{ or } 0 < k < n \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n.$$

Running Example. As before, $\mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = 2\theta_H\theta_T$ for all $\hat{n} \geq 2$. If we also take into account that $\{H,T\}[(3,1)] = \{H,T\}$, we get:

$$\underline{P}^{\hat{n}}_{\mathrm{SC},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \underline{P}^{\hat{n}}_{\mathrm{SC},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \min_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} \mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} 2\theta_H\theta_T = 0$$

and

$$\overline{P}^{\hat{n}}_{\mathrm{SC},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \overline{P}^{\hat{n}}_{\mathrm{SC},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \max_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} \mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\Sigma_{\{H,T\}}} 2\theta_H\theta_T = \frac{1}{2}.$$

Because all categories are observed for the count vector $(3,1)$—meaning that $\{H,T\}[(3,1)] = \{H,T\}$—we find the same inferences as for the vacuous inference system. $\Diamond$

Interestingly, the coherent inference system Φ_{SC} also satisfies both representation insensitivity and specificity.

Theorem 20. *The skeptically cautious inference system Φ_{SC} is coherent, representation insensitive and specific: $\Phi_{\mathrm{SC}} \in \mathbb{C}_{\mathrm{rs}}$.*

12. The IDMM Inference Systems

Imprecise Dirichlet Models—or IDMs, for short—are a family of parametric inference models introduced by Walley (1996) as conveniently chosen sets of *Dirichlet densities* $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ with constant prior weight s:

$$\{\mathrm{di}_A(\cdot|\boldsymbol{\alpha}) \colon \boldsymbol{\alpha} \in \mathrm{K}^s_A\}, \text{ with } \mathrm{K}^s_A := \left\{\boldsymbol{\alpha} \in \mathbb{R}^A_{>0} \colon \alpha_A = s\right\} = \{s\boldsymbol{t} \colon \boldsymbol{t} \in \mathrm{int}(\Sigma_A)\}, \tag{53}$$

for any value of the (so-called) hyperparameter $s \in \mathbb{R}_{>0}$ and any category set A. The Dirichlet densities $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ are defined on $\mathrm{int}(\Sigma_A)$; see Appendix C for an explicit definition and extensive discussion.

These IDMs generalise the Imprecise Beta models introduced earlier by Walley (1991). In a later paper, Walley and Bernard (1999) focussed on a closely related family of predictive inference models, which they called the Imprecise Dirichlet Multinomial Models—or IDMMs, for short.[33] We refer to these papers, and to a more recent overview paper by Bernard (2005) for extensive motivating discussion of the IDM(M)s, their inferences and properties. For precise Dirichlet models and their expectations, and the related Dirichlet multinomial models, we have gathered in Appendix C the most important facts, properties and results, necessary for a proper understanding of our present discussion of the IDM(M)s in the context of inference systems.

One of the reasons Walley (1996) had for suggesting the IDM as a reasonable model is precisely that it satisfies the pooling[34] invariance properties we discussed in Section 4.1. This is also discussed with more emphasis by Walley and Bernard (1999) and Bernard (2005), but we know of no detailed and explicit formulations of these properties in the literature, and the proofs we have seen are fairly sketchy. Bernard (1997, 2005) also suggests that the IDM and the underlying precise Dirichlet models satisfy a so-called 'specificity' property, which we have tried to translate to the present context of predictive inference in Section 4.5.

In the present section, we use the ideas behind Walley and Bernard's IDM(M)s to construct an interesting family of coherent inference systems, and we give a detailed and formal proof in Appendix E of the fact that these inference systems are indeed representation insensitive and specific. Interestingly, we shall need a slightly modified version of Walley's IDM(M) to make things work. The reason for this is that Walley's original version, as described by Expression (53), has a number of less desirable properties, that seem to have been either unknown to, or ignored by, Walley and Bernard. We describe these shortcomings in some detail in Appendix D. For our present purposes, it suffices to mention that, contrary to what is often claimed, and in contradistinction with our new version, inferences using the original version of the IDM(M) do not necessarily become more conservative (or less committal) as the hyperparameter s increases.

In our version, rather than using the hyperparameter sets K_A^s, we consider the sets

$$\Delta_A^s \coloneqq \left\{\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A \colon \alpha_A < s\right\} \text{ for all } s \in \mathbb{R}_{>0}.$$

Observe that

$$\Delta_A^s = \left\{s'\boldsymbol{t} \colon s' \in \mathbb{R}_{>0}, s' < s \text{ and } \boldsymbol{t} \in \operatorname{int}(\Sigma_A)\right\} = \bigcup_{0<s'<s} \mathrm{K}_A^{s'}.$$

For any $s \in \mathbb{R}_{>0}$, and any category set A, we now consider the following set of polynomials p, with positive Dirichlet expectation $\mathrm{Di}_A(p|\boldsymbol{\alpha})$ for all hyperparameters $\boldsymbol{\alpha} \in \Delta_A^s$:

$$\mathcal{H}_{\mathrm{IDM},A}^s \coloneqq \left\{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{\alpha} \in \Delta_A^s)\, \mathrm{Di}_A(p|\boldsymbol{\alpha}) > 0\right\}.$$

We shall see further on in Theorem 21 that this set is Bernstein coherent. We call the inference system Φ_{IDM}^s, defined by:

$$\Phi_{\mathrm{IDM}}^s(A) \coloneqq \mathcal{H}_{\mathrm{IDM},A}^s \text{ for all category sets } A,$$

33. In the later paper, Walley and Bernard (1999) clearly distinguish in name between the parametric IDMs and the predictive IDMMs, while in the earlier paper by Walley (1996), both types of models are referred to as IDMs.

34. Walley uses the term 'representation invariance' rather than 'pooling invariance'.

the *IDMM inference system* with hyperparameter $s > 0$. The corresponding updated models are, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$, given by:

$$\mathcal{H}^s_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}} = \{p \in \mathcal{V}(A)\colon (\forall \boldsymbol{\alpha} \in \Delta^s_A)\,\mathrm{Di}_A(p|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0\} \tag{54}$$

and

$$\underline{H}^s_{\mathrm{IDM},A}(p|\check{\boldsymbol{m}}) = \inf_{\boldsymbol{\alpha}\in\Delta^s_A} \mathrm{Di}_A(p|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) \text{ for all } p \in \mathcal{V}(A). \tag{55}$$

Using these expressions, the predictive models for the IDMM inference system are straightforward to find; it suffices to apply Equations (21) and (23). For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^{s,\hat{n}}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A^{\hat{n}})\colon (\forall \boldsymbol{\alpha} \in \Delta^s_A)\,\mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0\right\}, \tag{56}$$

and

$$\underline{P}^{s,\hat{n}}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}}) = \inf_{\boldsymbol{\alpha}\in\Delta^s_A} \mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}), \tag{57}$$

where, using the notations introduced in Appendix C:

$$\begin{aligned}\mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) &= \mathrm{DiMn}^{n}_A(\mathrm{Hy}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha})\\ &= \sum_{\hat{\boldsymbol{m}}\in\mathcal{N}^{\hat{n}}_A} \mathrm{Hy}^{\hat{n}}_A(f|\hat{\boldsymbol{m}}) \frac{1}{(\check{m}_A + \alpha_A)^{(\hat{n})}} \binom{\hat{n}}{\hat{\boldsymbol{m}}} \prod_{x\in A} (\check{m}_x + \alpha_x)^{(\hat{m}_x)}.\end{aligned} \tag{58}$$

In general, these expressions seem forbidding, but the immediate prediction models are manageable enough. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A)\colon f > -\frac{1}{s}\sum_{x\in A} f(x)\check{m}_x\right\}, \tag{59}$$

$$\underline{P}^{s,1}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}}) = \frac{1}{\check{m}_A + s}\sum_{x\in A} f(x)\check{m}_x + \frac{s}{\check{m}_A + s}\min f \text{ for all } f \in \mathcal{L}(A), \tag{60}$$

and

$$\varphi^s_{\mathrm{IDM}}(n,k) = \frac{k}{n+s} \text{ for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n.$$

The corresponding imprecision function is given by:

$$\iota^s_{\mathrm{IDM}}(n,k) = \frac{s}{n+s} \text{ for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n,$$

and it is decreasing in its first and constant in its second argument, which implies that it satisfies Condition (35). This suggests that IDMM inference systems are conservative when little has been learned, and become more precise as more observations come in.

Running Example. As before, with $\mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta}) = 2\theta_H\theta_T$ for all $\hat{n} \geq 2$, we find that, using the results in Appendix C:

$$\mathrm{Di}_{\{H,T\}}\big(\mathrm{Mn}^{\hat{n}}_{\{H,T\}}(\mathbb{I}_{\widehat{HT}}|\boldsymbol{\theta})\big|\boldsymbol{\alpha}\big) = \frac{2\alpha_H\alpha_T}{(\alpha_H + \alpha_T)(\alpha_H + \alpha_T + 1)}.$$

It is then not very difficult to verify using Equation (57) that for any $0 < s$:

$$\underline{P}^{s,\hat{n}}_{\mathrm{IDM},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = 0 \text{ and } \overline{P}^{s,\hat{n}}_{\mathrm{IDM},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \frac{1}{2}\frac{s}{1+s}.$$

After observing the count vector $(3,1)$, we find after some manipulations that:

$$\underline{P}^{s,\hat{n}}_{\mathrm{IDM},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \inf_{0<\sigma<s} \frac{2(3+\sigma)}{(4+\sigma)(5+\sigma)} = \frac{2(3+s)}{(4+s)(5+s)},$$

and similarly:

$$\overline{P}^{s,\hat{n}}_{\mathrm{IDM},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \begin{cases} 6\dfrac{1+s}{(4+s)(5+s)} & \text{if } s \leq 2 \\ \dfrac{1}{2}\dfrac{4+s}{5+s} & \text{if } s \geq 2. \end{cases}$$

Observe that for infinitely large s, we recover the inferences for the vacuous system. $\lozenge$

Interestingly, the immediate prediction models for our version of the IDMM inference system coincide with those of Walley's original version. Hence, in the many practical applications that are concerned with immediate prediction only, both approaches yield identical results.

The IDMM inference systems constitute an uncountably infinite family of coherent inference systems, each of which satisfies the representation insensitivity and specificity requirements.

Theorem 21. *For any $s \in \mathbb{R}_{>0}$, the IDMM inference system Φ^s_{IDM} is coherent, representation insensitive and specific: $\Phi^s_{\mathrm{IDM}} \in \mathbb{C}_{\mathrm{rs}}$.*

Since $\mathbb{C}_{\mathrm{rs}}$ is closed under non-empty infima, the infimum $\Phi^\infty_{\mathrm{IDM}}$ of all Φ^s_{IDM}, $s > 0$ is still coherent, representation insensitive and specific, and more conservative than any of the IDMM inference systems. It is given by:

$$\Phi^\infty_{\mathrm{IDM}}(A) = \mathcal{V}^{+++}(A) := \left\{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{\alpha} \in \mathbb{R}^A_{>0})\, \mathrm{Di}_A(p|\boldsymbol{\alpha}) > 0\right\},$$

and although this set generally strictly includes the sets $\mathcal{V}^+(A)$ and $\mathcal{V}^{++}(A)$, the associated immediate prediction models and predictive lower previsions can be shown to coincide with the ones for the vacuous and nearly vacuous inference systems.

13. The Skeptical IDMM Inference Systems

We now combine the ideas in the previous two sections: we suppose that our subject uses the following system for making inferences based on a sequence of $\check{n} > 0$ observations with count vector $\check{\boldsymbol{m}}$, in a category set A. As before in Section 11, he is skeptical in that he believes that in the future, he will only observe categories that he has seen previously, so only categories in the set $A[\check{\boldsymbol{m}}]$. But rather than being cautious in having completely vacuous beliefs about which of these already observed categories will be observed in the future, he uses an IDMM-like inference for them, as described in Section 12.

It turns out this can be done quite simply by replacing, in the characterisation (45) of the sets $\mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$ of the skeptically cautious inference system, the nearly vacuous models

$\mathcal{V}^{++}(K)$ by the appropriate IDMM models $\mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$. So we define, for any category set A, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ and any $s \in \mathbb{R}_{>0}$, the following set of polynomials:

$$\mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}} := \left\{p \in \mathcal{V}(A) \setminus \{0\} \colon (\forall K \in \min S_{A,\boldsymbol{m}}(p)) p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})\right\}, \tag{61}$$

where we recall that if $K \in \min S_{A,\boldsymbol{m}}(p)$, then $A[\boldsymbol{m}] \subseteq K$ and therefore $K[r_K(\boldsymbol{m})] = A[\boldsymbol{m}] \cap K = A[\boldsymbol{m}]$, so $\boldsymbol{m}$ and $r_K(\boldsymbol{m})$ are essentially the same count vectors. We also let $\mathcal{H}^s_{\mathrm{SI},A} := \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$ for $\boldsymbol{m} = 0$, or in other words:

$$\mathcal{H}^s_{\mathrm{SI},A} := \left\{p \in \mathcal{V}(A) \setminus \{0\} \colon (\forall K \in \min S_{A,\mathbf{0}}(p)) p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\right\},$$

where, again, $S_{A,\mathbf{0}}(p) = \left\{\emptyset \neq K \subseteq A \colon p_{|\Sigma_K} \neq 0\right\}$. In the remainder of this section, we show that the sets of polynomials $\mathcal{H}^s_{\mathrm{SI},A}$ indeed lead to the definition of a reasonable and potentially useful type of inference system. We begin with coherence.

Proposition 22. *$\mathcal{H}^s_{\mathrm{SI},A}$ is a Bernstein coherent set of polynomials on Σ_A.*

This shows that the inference system Φ^s_{SI}, given by $\Phi^s_{\mathrm{SI}}(A) := \mathcal{H}^s_{\mathrm{SI},A}$ for all category sets A, is coherent. We call Φ^s_{SI} the *skeptical IDMM inference system* with hyperparameter s.

We now want to find out how updating works in this inference system. The following proposition should not really come as a surprise.

Proposition 23. *For any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$: $\mathcal{H}^s_{\mathrm{SI},A}\rfloor\boldsymbol{m} = \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$.*

By combining this with Equation (21), we obtain the following—again, rather involved—predictive sets of desirable gambles for the skeptical IDMM inference systems. For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^{s,\hat{n}}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A^{\hat{n}}) \colon \operatorname{Mn}^{\hat{n}}_A(f) \in \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}}\right\}. \tag{62}$$

Although the expressions for $\mathcal{H}^s_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}}$ are rather abstract, this is not the case for the corresponding lower previsions. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{H}^s_{\mathrm{SI},A}(p) = \min_{x \in A} p(\boldsymbol{\theta}^\circ_x) \text{ for all } p \in \mathcal{V}(A) \tag{63}$$

and

$$\underline{H}^s_{\mathrm{SI},A}(p|\check{\boldsymbol{m}}) = \inf_{\boldsymbol{\alpha} \in \Delta^s_{A[\check{\boldsymbol{m}}]}} \operatorname{Di}_{A[\check{\boldsymbol{m}}]}(p_{|\Sigma_{A[\check{\boldsymbol{m}}]}}|r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}}) + \boldsymbol{\alpha}) \tag{64}$$

$$= \underline{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}(p_{|\Sigma_{A[\check{\boldsymbol{m}}]}}|r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}})) \text{ for all } p \in \mathcal{V}(A). \tag{65}$$

Combining this with Equation (23), we immediately obtain the following predictive lower previsions for the skeptical IDMM inference systems. For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^{s,\hat{n}}_{\mathrm{SI},A}(f) = \min_{x \in A} f(x, x, \dots, x) \text{ for all } f \in \mathcal{L}(A^{\hat{n}})$$

and

$$\underline{P}^{s,\hat{n}}_{\mathrm{SI},A}(f|\check{\boldsymbol{m}}) = \inf_{\boldsymbol{\alpha} \in \Delta^s_{A[\check{\boldsymbol{m}}]}} \operatorname{Di}_{A[\check{\boldsymbol{m}}]}(\operatorname{Mn}^{\hat{n}}_{A[\check{\boldsymbol{m}}]}(f_{|A[\check{\boldsymbol{m}}]^{\hat{n}}})|r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}}) + \boldsymbol{\alpha})$$

$$= \underline{P}^{s,\hat{n}}_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}(f_{|A[\check{\boldsymbol{m}}]^{\hat{n}}}|r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}})) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}). \tag{66}$$

The immediate prediction models of the skeptical IDMM inference systems are surprisingly more manageable:

$$\mathcal{D}^{s,1}_{\mathrm{SI},A} = \mathcal{L}_{>0}(A) \text{ and } \underline{P}^{s,1}_{\mathrm{SI},A}(f) = \min f \text{ for all } f \in \mathcal{L}(A)$$

and, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\mathcal{D}^{s,1}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}} = \left\{ f \in \mathcal{L}(A)\colon f_{|A[\check{\boldsymbol{m}}]} > -\frac{1}{s} \sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\check{m}_x \right\} \cup \mathcal{L}_{>0}(A) \tag{67}$$

$$\underline{P}^{s,1}_{\mathrm{SI},A}(f|\check{\boldsymbol{m}}) = \frac{1}{\check{m}_A + s} \sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\check{m}_x + \frac{s}{\check{m}_A + s} \min_{x\in A[\check{\boldsymbol{m}}]} f(x) \text{ for all } f \in \mathcal{L}(A). \tag{68}$$

The lower probability function is given by:

$$\varphi^s_{\mathrm{SI}}(n,k) = \begin{cases} \frac{k}{n+s} & \text{if } k < n \text{ or } n = 0 \\ 1 & \text{if } k = n > 0 \end{cases} \qquad \text{for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n,$$

and the corresponding imprecision function by:

$$\iota^s_{\mathrm{SI}}(n,k) = \begin{cases} \frac{s}{n+s} & \text{if } n = 0 \text{ or } 0 < k < n \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } n,k \in \mathbb{N}_0 \text{ such that } k \leq n.$$

When we consider the case $n > 0$, we see that $\iota^s_{\mathrm{SI}}(n,n) = 0$ but $\iota^s_{\mathrm{SI}}(n+1,n) = \frac{s}{n+1+s} > 0$, so this imprecision function does not satisfy Condition (35).

Running Example. Because $\{H,T\}[(3,1)] = \{H,T\}$, we infer from Equation (66) that the inferences about the event $\widehat{HT}$ are the same as for the IDMM inference systems. $\lozenge$

All the coherent inference systems Φ^s_{SI} also satisfy both representation insensitivity and specificity.

Theorem 24. *For each $s \in \mathbb{R}_{>0}$, the corresponding skeptical IDMM inference system is coherent, representation insensitive, and specific: $\Phi^s_{\mathrm{SI}} \in \mathbb{C}_{\mathrm{rs}}$.*

Since $\mathbb{C}_{\mathrm{rs}}$ is closed under non-empty infima, the infimum $\Phi^\infty_{\mathrm{SI}}$ of all Φ^s_{SI}, $s > 0$ is still coherent, representation insensitive and specific, and more conservative than any of the skeptical IDMM inference systems. It can be shown that the associated immediate prediction models and predictive lower previsions coincide with the ones for the skeptically cautious inference system.

14. The Haldane Inference System

We already know from our discussion of near-ignorance following Theorem 4 that no representation insensitive coherent inference system can be fully precise, as its immediate prediction models before observations have been made, must be completely vacuous. But we can ask ourselves whether there are representation insensitive (and specific) inference systems whose *posterior* predictive lower previsions become precise (linear) previsions. This is the problem we address in this section. We shall first construct such an inference system, and then show that this system is, in some definite sense, unique in having linear posterior predictive previsions.

We use the family of all IDMM inference systems Φ^s_{IDM}, $s \in \mathbb{R}_{>0}$, to define an inference system Φ_{H} that is more committal than any of them:

$$\Phi_{\mathrm{H}}(A) = \mathcal{H}_{\mathrm{H},A} \coloneqq \bigcup_{s\in\mathbb{R}_{>0}} \mathcal{H}^s_{\mathrm{IDM},A} = \bigcup_{s\in\mathbb{R}_{>0}} \Phi^s_{\mathrm{IDM}}(A) \text{ for all category sets } A.$$

We call this Φ_{H} the *Haldane inference system*, for reasons that will become clear further on in this section.

Theorem 25. *The Haldane inference system Φ_{H} is coherent, representation insensitive and specific: $\Phi_{\mathrm{H}} \in \mathbb{C}_{\mathrm{rs}}$.*

Due to its representation insensitivity, the Haldane system satisfies prior near-ignorance. This implies that before making any observation, its immediate prediction model is vacuous, and as far away from a precise probability model as possible. But we are about to show that, after making even a single observation, its inferences become precise-probabilistic: they coincide with the inferences generated by the Haldane (improper) prior.

To get there, we first take a look at the models involving sets of desirable gambles. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{H}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} = \{p \in \mathcal{V}(A)\colon (\exists s \in \mathbb{R}_{>0})(\forall \boldsymbol{\alpha} \in \Delta^s_A)\,\mathrm{Di}_A(p|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) > 0\} = \bigcup_{s\in\mathbb{R}_{>0}} \mathcal{H}^s_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}. \tag{69}$$

The corresponding predictive models are easily derived by applying Equation (21). For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\begin{aligned}\mathcal{D}^{\hat{n}}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} &= \Big\{f \in \mathcal{L}(A^{\hat{n}})\colon (\exists s \in \mathbb{R}_{>0})(\forall \boldsymbol{\alpha} \in \Delta^s_A)\,\mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) > 0\Big\}\\ &= \bigcup_{s\in\mathbb{R}_{>0}} \mathcal{D}^{s,\hat{n}}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}.\end{aligned} \tag{70}$$

The immediate prediction models are obtained by combining Equations (70) and (59). For any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\mathcal{D}^1_{\mathrm{H},A} = \mathcal{L}_{>0}(A) \text{ and } \mathcal{D}^1_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} = \Big\{f \in \mathcal{L}(A)\colon \sum_{x\in A} f(x)\check{m}_x > 0\Big\} \cup \mathcal{L}_{>0}(A).$$

It turns out that the expressions for the corresponding lower previsions are much more manageable. First of all, we find for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\underline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = \lim_{s\to+0} \inf_{\boldsymbol{\alpha}\in\Delta^s_A} \mathrm{Di}_A(p|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) = \lim_{s\to+0} \underline{H}^s_{\mathrm{IDM},A}(p|\check{\boldsymbol{m}}) \text{ for all } p \in \mathcal{V}(A). \tag{71}$$

In particular, for $\check{\boldsymbol{m}} = \mathbf{0}$, this simplifies to:

$$\underline{H}_{\mathrm{H},A}(p) = \min_{x\in A} p(\boldsymbol{\theta}^\circ_x) \text{ for all } p \in \mathcal{V}(A), \tag{72}$$

whereas for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$, we find *linear* previsions:[35]

$$\underline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = \overline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = H_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = \mathrm{Di}_A(p|\check{\boldsymbol{m}}) \text{ for all } p \in \mathcal{V}(A). \tag{73}$$

The corresponding predictive models are easily derived by applying Equation (23). For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\underline{P}^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \lim_{s\to+0} \inf_{\boldsymbol{\alpha}\in\Delta^s_A} \mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) = \lim_{s\to+0} \underline{P}^{s,\hat{n}}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}}) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}). \tag{74}$$

In particular, for $\check{\boldsymbol{m}} = \mathbf{0}$:

$$\underline{P}^{\hat{n}}_{\mathrm{H},A}(f) = \min_{x\in A} f(x, x, \ldots, x) \text{ for all } f \in \mathcal{L}(A^{\hat{n}}),$$

and for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \overline{P}^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = P^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \sum_{\boldsymbol{n}\in\mathcal{N}^{\hat{n}}_A} \mathrm{Hy}^{\hat{n}}_A(f|\boldsymbol{n}) \binom{\hat{n}}{\boldsymbol{n}} \frac{\prod_{x\in A} \check{m}_x^{(n_x)}}{\check{m}_A^{(\hat{n})}}. \tag{75}$$

For the immediate prediction models, we find that for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^{1}_{\mathrm{H},A}(f) = \min f \text{ and } P^{1}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \sum_{x\in A} f(x)\frac{\check{m}_x}{\check{m}_A} \text{ for all } f \in \mathcal{L}(A),$$

and the lower probability function is given by:

$$\varphi_{\mathrm{H}}(n,k) = \begin{cases} \frac{k}{n} & \text{if } n > 0 \\ 0 & \text{if } n = 0 \end{cases} \quad \text{for all } n, k \in \mathbb{N}_0 \text{ such that } k \leq n.$$

The corresponding imprecision function is given by:

$$\iota_{\mathrm{H}}(n,k) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n > 0 \end{cases} \quad \text{for all } n, k \in \mathbb{N}_0 \text{ such that } k \leq n,$$

and it satisfies Condition (35), which suggests that also the Haldane inference system displays—albeit in an extreme and not very interesting manner—the desirable behaviour mentioned in the Introduction: it is conservative when little has been learned, and it never become less precise as more observations come in.

Running Example. We can use Equation (74) and the results previously obtained for the IDMM inference systems to find that

$$\underline{P}^{\hat{n}}_{\mathrm{H},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = \overline{P}^{\hat{n}}_{\mathrm{H},\{H,T\}}(\mathbb{I}_{\widehat{HT}}) = 0 \text{ and } P^{\hat{n}}_{\mathrm{H},\{H,T\}}(\mathbb{I}_{\widehat{HT}}|(3,1)) = \frac{3}{10}.$$

We want to point out that the first equalities do not contradict the prior near-ignorance of the Haldane inference system, as that only pertains to immediate predictions: predictions about *single* future observations. ♢

35. The Dirichlet expectations $\mathrm{Di}_A(\cdot|\boldsymbol{\alpha})$ are strictly speaking defined for $\boldsymbol{\alpha} \in \mathbb{R}^A_{>0}$, but as we argue in Appendix C, they can be continuously extended to $\boldsymbol{\alpha}$ with some components zero, and the others strictly positive.

The precise posterior predictive previsions in Equation (75) are exactly the ones that would be found were we to formally apply Bayes's rule with a multinomial likelihood and *Haldane's improper prior* (Haldane, 1945; Jeffreys, 1998; Jaynes, 2003), whose 'density' is a function on $\mathrm{int}(\Sigma_A)$ proportional to $\prod_{x\in A}\theta_x^{-1}$. This, of course, is why we use Haldane's name for the inference system that produces them. Our argumentation shows that there is nothing wrong with these posterior predictive previsions, as they are based on coherent inferences. In fact, our analysis shows that *there is an infinity of* precise and proper *priors on the simplex* Σ_A *that, together with the multinomial likelihood, are coherent with these posterior predictive previsions:* every coherent prevision on $\mathcal{V}(A)$ that dominates the coherent lower prevision $\underline{H}_{\mathrm{H},A}$ on $\mathcal{V}(A)$.[36] For binomial parametric inferences under the Haldane prior, Walley (1991, Section 7.4.8) comes to a related conclusion in a completely different manner.

There is a simple argument to show that these Haldane posterior predictive previsions are the only precise ones that are compatible with representation insensitivity. Indeed, it can be shown that for any representation insensitive coherent inference system with precise posterior predictive previsions, the lower probability function must satisfy $\varphi(n,k)=k/n$ for $n>0$ and $0\leq k\leq n$,[37] and then it is straightforward to prove, using Bayes's Theorem to go from immediate prediction to more general predictive inference, that the posterior predictive previsions must be Haldane's.

15. Characterisation of the IDMM Immediate Predictions

The lower probability function $\varphi(n,k)$ for a representation insensitive coherent inference system gives the lower probability of observing a non-trivial event that has been observed k times before in n trials.

Now suppose that a subject specifies a single lower probability, namely the value of $\varphi(1,1)\in[0,1]$: the probability of observing something (again) that has been observed (once) in a single trial. Then we can ask ourselves what the most conservative consequences of such an assessment are, if we take representation insensitivity and specificity for granted. In other words, *what is the most conservative representation insensitive and specific coherent inference system that has (at least) this given value* $\varphi(1,1)$ *for its lower probability function?* This question makes sense because the representation insensitive and specific coherent inference systems constitute a complete meet-semilattice by Statement (25) and Theorems 8 and 12.[38]

Clearly, if $\varphi(1,1)=0$, this is the smallest representation insensitive and specific coherent inference system, which as we know from the discussion in Sections 9 and 10, must have the same immediate prediction models and predictive lower previsions as the (nearly) vacuous inference system. We consider the case that $0<\varphi(1,1)<1$,[39] or in other words, to use a

36. It is an immediate consequence of the F. Riesz Representation Theorem that each such coherent prevision is the restriction to polynomials of the expectation operator of some unique σ-additive probability measure on the Borel sets of Σ_A; see for instance the discussion by De Cooman and Miranda (2008a) and also footnote 2.

37. It suffices to exploit the additivity of precise probabilities and the symmetry implied by representation insensitivity; for an explicit proof, see the paper by De Cooman et al. (2009a, Thm. 7).

38. See the discussion near the end of Section 7.

39. We surmise, but do not prove here, that the most conservative representation insensitive and specific coherent inference system corresponding to $\varphi(1,1)=1$ might be the skeptically cautious one.

parametrisation that will turn out to be more convenient for our purposes, that:

$$\varphi(1,1)=\frac{1}{1+s}\text{ for some positive real number }s\coloneqq\frac{1}{\varphi(1,1)}-1. \tag{76}$$

Let us denote this most conservative inference system by Φ^s, and its lower probability function by φ^s, then by assumption $\varphi^s(1,1)\geq\frac{1}{1+s}$. It now follows from Theorem 14.L9 that $\varphi^s(n,n)\geq\frac{n}{n+s}$ for all $n\in\mathbb{N}_0$. But since for the IDMM inference system Φ^s_{IDM}, Equation (60) tells us that $\varphi^s_{\mathrm{IDM}}(n,n)=\frac{n}{n+s}$, and since by assumption $\varphi^s_{\mathrm{IDM}}(n,n)\geq\varphi^s(n,n)$, we conclude that:

$$\varphi^s(n,n)=\varphi^s_{\mathrm{IDM}}(n,n)=\frac{n}{n+s}\text{ for all }n\in\mathbb{N}_0. \tag{77}$$

It has been surmised (Bernard, 2007; De Cooman et al., 2009a) that the IDMM inference system with hyperparameter s could be the smallest, most conservative, representation insensitive and specific coherent inference system with a given value $\varphi(1,1)=\frac{1}{1+s}$. In fact, trying to prove this was what made us start research on the present paper. But this conjecture turns out to be false: apart from the lower bound (77) on the $\varphi(n,n)$, representation insensitivity and specificity impose no lower bounds on the $\varphi(n,k)$ for $k<n$. To see this, consider the inference system $\Phi^s_{\mathrm{MC}}\coloneqq\inf\{\Phi_{\mathrm{SC}},\Phi^s_{\mathrm{IDM}}\}$, which by Statement (25) and Theorems 8, 12, 20 and 21 is coherent, representation insensitive and specific: $\Phi^s_{\mathrm{MC}}\in\mathbb{C}_{\mathrm{rs}}$. Its lower probability function φ^s_{MC} satisfies:

$$\varphi^s_{\mathrm{MC}}(n,k)=\min\{\varphi_{\mathrm{SC}}(n,k),\varphi^s_{\mathrm{IDM}}(n,k)\}=\begin{cases}\min\{1,\frac{n}{n+s}\}=\frac{n}{n+s} & \text{if }k=n>0\\ \min\{0,\frac{k}{n+s}\}=0 & \text{otherwise,}\end{cases}$$

substantiating the claim we made above. See also Figure 1, where we have depicted lower (and upper) probability functions for the Haldane system Φ_{H}, the IDMM system Φ^s_{IDM}, Φ^s_{MC} and the inference system $\inf\{\Phi^{4s}_{\mathrm{SI}},\Phi^s_{\mathrm{IDM}}\}$. The latter three all share the same value $\frac{n}{n+s}$ for $\varphi(n,n)$, $n\geq 0$. We conjecture that Φ^s_{MC} could be the smallest, most conservative, representation insensitive and specific coherent inference system with a given value $\varphi(1,1)=\frac{1}{1+s}$, but offer no proof for this.

This means that if we want to characterise the IDMM inference systems in any way as the most conservative ones, we need to add, besides coherence, representation insensitivity and specificity, another requirement that is preserved under taking infima. One possible candidate for this, which we shall prove does the job and is inspired by Figure 1, is the following requirement.

Let us define the subject's *surprise* of an event as his supremum rate for betting on the opposite event, or in other words, his lower probability for the opposite event. This surprise is high—close to one—when the subject believes strongly that the event will not occur, and low—close to zero—when the subject has no strong beliefs that it will not occur.

This allows us to associate a so-called *surprise function* $\varsigma(n,k)\coloneqq\varphi(n,n-k)$ with a lower probability function, where $\varsigma(n,k)$ is the subject's surprise when observing a non-trivial event that has been observed k out of n times before.

It follows from Theorem 14.L5 that for any representation insensitive system, the surprise function is non-increasing in its second argument:

$$\Delta\varsigma(n,k)\coloneqq\varsigma(n,k+1)-\varsigma(n,k)=\varphi(n,n-k-1)-\varphi(n,n-k)\leq 0\text{ for }0\leq k\leq n-1.$$

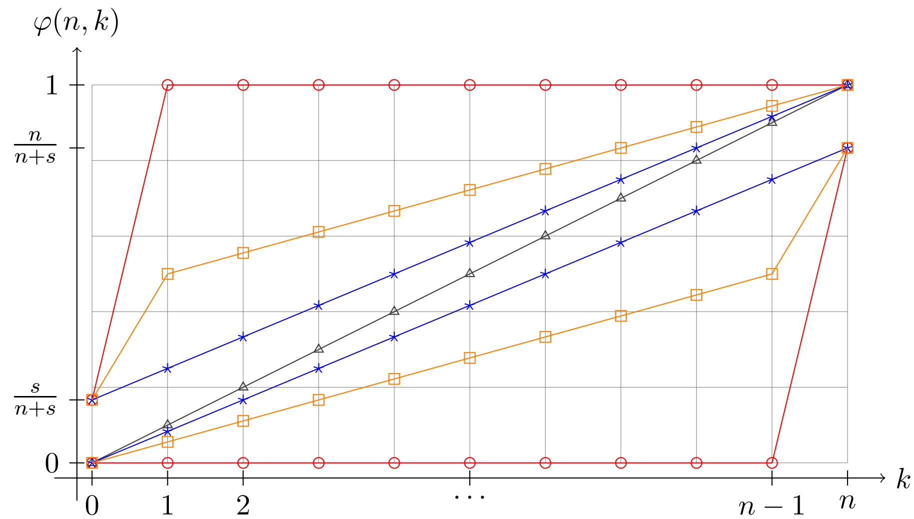

Figure 1: Lower and upper probability functions: φ_{H} for the Haldane system (dark grey, $\triangle$), φ^s_{IDM} for the IDMM system with hyperparameter s (blue, $\star$), $\min\{\varphi^{4s}_{\mathrm{SI}}, \varphi^s_{\mathrm{IDM}}\}$ (orange, $\square$) and $\varphi^s_{\mathrm{MC}} = \min\{\varphi_{\mathrm{SC}}, \varphi^s_{\mathrm{IDM}}\}$ (red, $\circ$). This specific plot was made for $n = 10$ and $s = 2$.

This is a fairly intuitive property: the more often an event has been observed before, the smaller the surprise is at seeing it again.

We shall say that a representation insensitive system *has concave surprise* if

$$\Delta^2\varsigma(n,k) \coloneqq \Delta\varsigma(n,k+1) - \Delta\varsigma(n,k) \leq 0 \text{ for } 0 \leq k \leq n-2,$$

where, of course, $\Delta^2\varsigma(n,k) = \varphi(n,n-k-2) - 2\varphi(n,n-k-1) + \varphi(n,n-k)$. It is not difficult to see that having concave surprise is preserved under taking non-empty infima of inference systems, so it makes sense to go looking for the smallest (most conservative) coherent representation insensitive and specific coherent inference system that has concave surprise, and satisfies some additional local assessments, such as (76).

Looking at Figure 1 makes us suspect that the IDMM inference system Φ^s_{IDM} might be this system, but again, we offer no proof for this conjecture. We can however provide a proof for the following, related but (probably) weaker, statement, which focusses on immediate prediction only:

Theorem 26. *The immediate prediction models $\underline{P}^1_A(\cdot|\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ for the smallest (most conservative) coherent representation insensitive and specific coherent inference system Φ that has concave surprise and satisfies* (76)*, coincide with the ones for the IDMM inference system Φ^s_{IDM} with hyperparameter s.*

16. Conclusion

We believe this is the first paper that tries to deal in a systematic fashion with principles for predictive inference under exchangeability using imprecise probability models. Two salient features of our approach are (i) that we consistently use coherent sets of desirable gambles as our uncertainty models of choice; and (ii) that our notion of an inference system allows us to derive a conservative predictive inference method combining both local predictive probability assessments and general inference principles.

The first feature is what allows us, in contradistinction with most other approaches in probability theory, to avoid problems with determining unique conditional models from unconditional ones when conditioning on events with (lower) probability zero. A set of polynomials $\mathcal{H}_A$ completely determines all prior and posterior predictive models $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}}$ and $\underline{P}_A^{\hat{n}}(\cdot|\check{\boldsymbol{m}})$, even when the (lower) prior probability $\underline{P}_A^{\check{n}}([\check{\boldsymbol{m}}]) = \underline{H}_A(\mathrm{B}_{A,\check{\boldsymbol{m}}})$ of observing the count vector $\check{\boldsymbol{m}}$ is zero. An approach using only lower previsions and probabilities would make this much more complicated and involved, if not impossible. Interestingly, we can provide a perfect illustration of this fact using the results in Sections 11, 13 and 14.[40] The three inference systems that are described there—the skeptically cautious, the skeptical IDMM and the Haldane systems—have, for any given category set A, three different sets of polynomials $\mathcal{H}_A$. Nevertheless, as we can gather from Equations (49), (63) and (72), they have the same lower prevision $\underline{H}_A$ and therefore the same prior predictive models $\underline{P}_A^{\hat{n}}$. And any count vector $\check{\boldsymbol{m}} \in \mathcal{N}_A$ has the same prior lower probability:

$$\underline{P}_A^{\check{n}}([\check{\boldsymbol{m}}]) = \underline{H}_A(\mathrm{B}_{A,\check{\boldsymbol{m}}}) = \min_{x\in A} \mathrm{B}_{A,\check{\boldsymbol{m}}}(\boldsymbol{\theta}_x^\circ) = 0.$$

This zero lower probability makes sure that the posterior lower previsions $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$ and the posterior predictive models $\underline{P}_A^{\hat{n}}(\cdot|\check{\boldsymbol{m}})$ are not uniquely determined by the prior lower prevision $\underline{H}_A$: we infer from Equations (49), (65) and (73) that they are indeed very different for these three types of inference systems. We fail to see how we could have come up with—let alone proved the necessary results for—these three systems relying only on lower prevision or credal set theory.

We can—and must—take this line of argumentation even further. By Theorem 4, any inference system that satisfies (prior) representation insensitivity has near-vacuous prior predictive models, and therefore, by time consistency and coherence [monotonicity], we see that its prior predictive lower previsions must satisfy $\underline{H}_A(\mathrm{B}_{A,\check{\boldsymbol{m}}}) = \underline{P}_A^{\check{n}}([\check{\boldsymbol{m}}]) = 0$ for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$ as well. This simply means that it is *impossible* in a (prior) representation insensitive coherent inference system for the lower prevision $\underline{H}_A$ to uniquely determine the conditional lower previsions $\underline{H}_A(\cdot|\check{\boldsymbol{m}})$. And therefore any systematic way of dealing with such inference systems must be able to resolve—or deal with—this non-unicity in some way. We believe our approach involving coherent sets of desirable gambles is one of the mathematically more elegant ways of doing this.

The second feature has allowed us, as an example, to characterise the IDMM immediate predictions as the most conservative ones satisfying a number of inference principles. The approach we follow can—at least in principle—also be used for other types of inference systems and other inference principles. The key requirement for an inference principle to make it amenable to our approach is that, when formulated as a property of an inference system, it should be preserved under taking arbitrary non-empty infima. The three inference

40. Something similarly 'dramatic' happens in Sections 9 and 10: the inference systems there have the same immediate prediction models and the same (predictive) lower previsions, but one is specific and the other is not.

principles that we have been considering above—representation insensitivity, specificity and having concave surprise—have this property, but there is nothing that prevents our analysis and approach from being extended to any other inference principle that has it too. The only complications we see, at this point, are of a technical mathematical nature. The reader will no doubt have noticed that our proofs for the results in the later sections are quite involved and technical, and rely quite heavily on properties of polynomials on a simplex. We feel that in the present paper we have made some headway into this mathematical territory, for instance with our new discussion about the Bernstein positivity of polynomials near Proposition 28 in Appendix B. In the Conclusions of a paper by De Cooman and Quaeghebeur (2012), a characterisation of Bernstein positivity was mentioned as an open problem with interesting practical applications in doing inference—natural extension—under exchangeability. But much remains open for further exploration, and a more determined study of the mathematical structure and properties of such polynomials would certainly help in alleviating the technical difficulties of working with inference principles in inference systems.

While this paper has only just opened up what we feel to be an interesting line of research into the foundations of predictive inference, it nevertheless has provided answers to a number of—if not all—open problems formulated in the Conclusions of an earlier paper by De Cooman et al. (2009a), who tried to deal with representation insensitivity in immediate prediction. As a first example: it was asked there whether there are representation insensitive coherent inference systems whose lower probability functions are not additive in the second argument? It suffices to look at Figure 1 to see that the answer is, clearly, yes. Another question was: are there representation insensitive coherent inference systems that are not mixing predictive systems?[41] It follows from Equation (68) that the answer is yes: each of the skeptical IDMM inference systems provides an example. Finally, we can use the infimum Φ^{s}_{MC} of the skeptically cautious inference system Φ_{SC} and an IDMM inference system Φ^{s}_{IDM}, mentioned briefly in Section 15, to answer two more questions. Are there representation insensitive coherent inference systems for which the inequality in Theorem 14.L6 is strict? And are there representation insensitive coherent inference systems whose behaviour on gambles is not completely determined by their lower probability function? The inference system Φ^{s}_{MC} provides a positive answer to both questions.

Most of the inference systems mentioned above, apart from the IDMM and the Haldane systems, appear here for the first time. Some of them may appear contrived and perhaps even artificial, but we have found them to be most useful in constructing (counter)examples, shaping intuition, and building new models, as Figure 1 and the argumentation above clearly indicate. We might also wonder whether there are other representation insensitive and/or specific coherent inference systems, which cannot be produced as appropriately chosen infima of the examples we have introduced here. We suggest, as candidates for further consideration, the inference systems that can be derived using Walley's (1997) bounded derivative model, and inference systems that can be constructed using sets of infinitely divisible distributions, as recently proposed by Mangili and Benavoli (2013). The framework provided here, as

41. Loosely speaking: that cannot be written as a (specific kind of) convex mixture of the Haldane inference system and an IDMM inference system; see the paper by De Cooman et al. (2009a, Section 5) for more information.

well as the simple characterisation results of Theorems 7 and 11, should be quite useful in addressing this and similar problems.

To end, we want to draw attention once again to a simple and direct, but quite appealing, consequence of our argumentation in Section 14: there is an infinity of *precise and proper* priors that, together with the multinomial likelihood, are coherent with the Haldane posterior predictive previsions. *So, there is no need for improper priors to 'justify' these posteriors, as there are proper priors that will do the job perfectly well.* This (precise-)probabilistic conclusion follows easily when looking at the problem using the more general and powerful language of imprecise probabilities. Moreover, we have seen that properties such as representation insensitivity cannot be satisfied by precise probabilistic models. Finally, the entire framework of conservative predictive inference using inference principles would be impossible to develop within the more limitative context of precise probabilities. This shows that there are distinct advantages to using imprecise probability models for dealing with predictive inference.

Acknowledgements

Gert de Cooman's research was partially funded through project number 3G012512 of the Research Foundation Flanders (FWO). Jasper De Bock is a PhD Fellow of the Research Foundation Flanders and wishes to acknowledge its financial support. Marcio Diniz was supported by FAPESP (São Paulo Research Foundation), under the project 2012/14764-0 and wishes to thank the SYSTeMS Research Group at Ghent University for its hospitality and support during his sabbatical visit there. The authors would like to thank three anonymous reviewers for their many insightful comments and suggestions aimed at making this paper easier to read and cleaning up misunderstandings. A special thank you also to the great Arthur Van Camp for his enthusiasm in everything and, in particular, in helping us check little examples.

Appendix A. Notation

In this appendix, we provide a list of the most commonly used and most important notation, and where it is defined or first introduced.

notation	meaning	introduced where
A, B, C, D	category sets, events	Section 1
$\mathbb{I}_B$	indicator of an event B	Section 2.2
X, X_n	variable, variable at time n	Section 1
$\check{n}$	number of already observed variables	Section 1
$\hat{n}$	number of to be observed variables	Section 1
$\mathrm{posi}(\mathcal{A})$	cone generated by $\mathcal{A}$	Equation (1)
$\mathcal{L}(A)$	set of all gambles on A	Section 2.2
$\mathcal{L}_{>0}(A)$	set of all positive gambles on A	Section 2.2
$\mathcal{L}_{\leq 0}(A)$	set of all non-positive gambles on A	Section 2.2
$\check{\boldsymbol{x}}$	observed sample	Section 3
$\check{\boldsymbol{m}}$	observed count vector	Section 5.4
$\mathcal{D}_A^{\hat{n}}$	prior predictive set of desirable gambles for category set A and $\hat{n}$ future observations	Section 3

$\mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{x}}$, $\mathcal{D}^{\hat{n}}_A\rfloor\check{\boldsymbol{m}}$	posterior predictive set of desirable gambles	Section 3
$\underline{P}^{\hat{n}}_A(\cdot)$	prior predictive lower prevision	Section 3
$\underline{P}^{\hat{n}}_A(\cdot\vert\check{\boldsymbol{x}})$, $\underline{P}^{\hat{n}}_A(\cdot\vert\check{\boldsymbol{m}})$	posterior predictive lower prevision	Section 3
ρ	pooling map or relabelling map	Sections 4.1&4.4
λ	renaming bijection	Section 4.2
ϖ	category permutation	Sections 4.3
$\check{\boldsymbol{x}}\downarrow_B$	sample with observations outside B eliminated	Section 4.5
$\boldsymbol{T}$	counting map	Equation (8)
$\mathcal{N}^n_A$	set of count vectors for n observations	Equation (9)
$\mathcal{N}_A$, $\mathcal{N}_A\cup\{\boldsymbol{0}\}$	set of all count vectors, with zero	Section 5.3
$\mathrm{Hy}^n_A(\cdot\vert\boldsymbol{m})$	hypergeometric expectation operator	Equation (10)
$\nu(\boldsymbol{m})$	multinomial coefficient with count vector $\boldsymbol{m}$	Equation (11)
$\mathrm{Mn}^n_A(\cdot\vert\boldsymbol{\theta})$	multinomial expectation operator	Equation (13)
Σ_A	simplex of all probability mass functions on A	Equation (12)
θ_B	sum of components θ_x of $\boldsymbol{\theta}$ over $x\in B$	Equation (12)
$\mathrm{B}_{A,\boldsymbol{m}}$	Bernstein basis polynomial	Equation (15)
$\mathcal{V}^n(A)$	set of polynomials of degree up to n on Σ_A	Section 5.3
$\mathcal{V}(A)$	set of all polynomials on Σ_A	Section 5.3
$\mathcal{V}^+(A)$	set of Bernstein positive polynomials on Σ_A	Section 5.3
$\mathcal{V}^{++}(A)$	set of polynomials on Σ_A that are positive on $\mathrm{int}(\Sigma_A)$	Section 10
$\mathcal{H}_A$	representing set of polynomials	Theorem 5
$\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$	updated representing set of polynomials	Equation (19)
$\underline{H}_A$	lower prevision induced by $\mathcal{H}_A$	Equation (17)
$\underline{H}_A(\cdot\vert\check{\boldsymbol{m}})$	lower prevision induced by $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$	Equation (20)
$\mathbb{F}$	set of all category sets	Definition 6
Φ	inference system	Definition 6
$\mathbb{C}$	set of all coherent inference systems	Equation (24)
$\mathbb{C}_{\mathrm{rs}}$	set of coherent inference systems that are representation insensitive and specific	Theorem 12
R_ρ	extended relabelling map	Equation (26)
r_B	restriction map	Equation (27)
i_A	injection map	Equation (28)
$\mathrm{I}^r_{B,A}$	extended injection map	Equation (29)
S_A	sampling expectation	Section 8
φ	lower probability function	Equation (33)
ι	imprecision function	Equation (34)
ς	surprise function	Section 15
subscript $_{\mathrm{V}}$	related to vacuous inference system	Section 9
subscript $_{\mathrm{NV}}$	related to nearly vacuous inference system	Section 10
subscript $_{\mathrm{SC}}$	related to skeptically cautious inference system	Section 11
subscript $_{\mathrm{IDM}}$	related to IDMM inference systems	Section 12
subscript $_{\mathrm{SI}}$	related to skeptical IDMM inference systems	Section 13
subscript $_{\mathrm{H}}$	related to Haldane inference system	Section 14
subscript $_{\mathrm{OI}}$	related to original IDMM inference systems	Appendix D

$A[\check{\boldsymbol{m}}]$	categories in A already observed	Equation (42)
$\mathcal{V}^{++}_{[\boldsymbol{m}]}(A)$	set of polynomials on Σ_A that are positive on $\operatorname{int}(\Sigma_{A[\boldsymbol{m}]})$	Section 11
$\operatorname{di}_A(\cdot\vert\boldsymbol{\alpha})$	Dirichlet density	Appendix C
$\operatorname{Di}_A(\cdot\vert\boldsymbol{\alpha})$	Dirichlet expectation operator	Appendix C
$\operatorname{DiMn}^n_A(\cdot\vert\boldsymbol{\alpha})$	Dirichlet multinomial expectation operator	Appendix C
b^n_p	expansion of polynomial p in Bernstein basis of degree n	Appendix B

Appendix B. Multivariate Bernstein Basis Polynomials

With any $n \geq 0$ and $\boldsymbol{m} \in \mathcal{N}^n_A$ there corresponds a (multivariate) *Bernstein basis polynomial* of degree n on Σ_A, given by $\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) := \nu(\boldsymbol{m}) \prod_{x\in A} \theta_x^{m_x}$, $\boldsymbol{\theta} \in \Sigma_A$. These polynomials have a number of very interesting properties (see for instance Prautzsch, Boehm, & Paluszny, 2002, Chapters 10 and 11), which we list here:

BB1. The set $\{\mathrm{B}_{A,\boldsymbol{m}} : \boldsymbol{m} \in \mathcal{N}^n_A\}$ of all Bernstein basis polynomials of fixed degree n is linearly independent: if $\sum_{\boldsymbol{m}\in\mathcal{N}^n_A} \lambda_{\boldsymbol{m}} \mathrm{B}_{A,\boldsymbol{m}} = 0$, then $\lambda_{\boldsymbol{m}} = 0$ for all $\boldsymbol{m}$ in $\mathcal{N}^n_A$.

BB2. The set $\{\mathrm{B}_{A,\boldsymbol{m}} : \boldsymbol{m} \in \mathcal{N}^n_A\}$ of all Bernstein basis polynomials of fixed degree n forms a *partition of unity*: $\sum_{\boldsymbol{m}\in\mathcal{N}^n_A} \mathrm{B}_{A,\boldsymbol{m}} = 1$.

BB3. All Bernstein basis polynomials are non-negative, and strictly positive on the interior $\operatorname{int}(\Sigma_A)$ of Σ_A.

BB4. The set $\{\mathrm{B}_{A,\boldsymbol{m}} : \boldsymbol{m} \in \mathcal{N}^n_A\}$ of all Bernstein basis polynomials of fixed degree n forms a basis for the linear space of all polynomials whose degree is at most n.

Property BB4 follows from BB1 and BB2.[42] It follows from BB4 that:

BB5. Any polynomial p has a unique expansion in terms of the Bernstein basis polynomials—also called *Bernstein expansion*—of fixed degree $n \geq \deg(p)$,

or in other words, there is a unique count gamble b^n_p on $\mathcal{N}^n_A$ such that:

$$p(\boldsymbol{\theta}) = \sum_{\boldsymbol{m}\in\mathcal{N}^n_A} b^n_p(\boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Sigma_A. \tag{78}$$

This tells us [also use BB2 and BB3] that each $p(\boldsymbol{\theta})$ is a convex combination of the Bernstein coefficients $b^n_p(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{N}^n_A$ whence:

$$\min b^n_p \leq \min p \leq p(\boldsymbol{\theta}) \leq \max p \leq \max b^n_p \text{ for all } \boldsymbol{\theta} \in \Sigma_A. \tag{79}$$

The following proposition adds more detail to this picture.

42. To see how: clearly all polynomials are by definition linear combinations of Bernstein basis polynomials, of possibly different degrees. For each of the terms, use BB2 to raise the degree to a common higher degree n—multiply it by an appropriate version of 1. This shows that the Bernstein basis polynomials of fixed degree n are generating for all polynomials of lower degrees. They are also independent by BB1.

Proposition 27. *For any polynomial p on Σ_A:*

$$\lim_{\substack{n\to+\infty\\ n\geq\deg(p)}}[\min b_p^n, \max b_p^n] = [\min p, \max p] = p(\Sigma_A).$$

Proof of Proposition 27. Since b_p^n converges uniformly to the polynomial p as $n \to +\infty$ (Trump & Prautzsch, 1996), in the sense that

$$\lim_{\substack{n\to+\infty\\ n\geq\deg(p)}} \max_{\boldsymbol{\mu}\in\mathcal{N}_A^n} \left| p\left(\frac{\boldsymbol{\mu}}{n}\right) - b_p^n(\boldsymbol{\mu})\right| = 0,$$

we find that

$$\begin{aligned}
\lim_{\substack{n\to+\infty\\ n\geq\deg(p)}} \min b_p^n - \min p &= \lim_{\substack{n\to+\infty\\ n\geq\deg(p)}} \min_{\boldsymbol{\mu}\in\mathcal{N}_A^n} \left[b_p^n(\boldsymbol{\mu}) - \min p\right] \\
&\geq \lim_{\substack{n\to+\infty\\ n\geq\deg(p)}} \min_{\boldsymbol{\mu}\in\mathcal{N}_A^n} \left[b_p^n(\boldsymbol{\mu}) - p\left(\frac{\boldsymbol{\mu}}{n}\right)\right] \\
&\geq - \lim_{\substack{n\to+\infty\\ n\geq\deg(p)}} \max_{\boldsymbol{\mu}\in\mathcal{N}_A^n} \left| p\left(\frac{\boldsymbol{\mu}}{n}\right) - b_p^n(\boldsymbol{\mu})\right| = 0,
\end{aligned}$$

and therefore $\lim_{n\to+\infty, n\geq\deg(p)} \min b_p^n \geq \min p$. Furthermore, by Statement (79), we see that $\lim_{n\to+\infty, n\geq\deg(p)} \min b_p^n \leq \min p$. Hence indeed $\lim_{n\to+\infty, n\geq\deg(p)} \min b_p^n = \min p$. The proof for the other equality is completely analogous. □

Using the above results, we can prove a number of useful relations between the Bernstein positivity of a polynomial and its positivity on (the interior of) the simplex. They are related to a property first proved by Hausdorff in the univariate case (Hausdorff, 1923, p. 124).

Proposition 28. *Let p be any polynomial on Σ_A. Consider the following statements:*

(i) $(\forall\boldsymbol{\theta}\in\Sigma_A)p(\boldsymbol{\theta}) > 0$;

(ii) $p \in \mathcal{V}^+(A)$, *meaning that there is some* $n \geq \deg(p)$ *such that* $b_p^n > 0$;

(iii) $p \in \mathcal{V}^{++}(A)$, *meaning that* $(\forall\boldsymbol{\theta}\in\operatorname{int}(\Sigma_A))p(\boldsymbol{\theta}) > 0$;

(iv) $(\forall\boldsymbol{\theta}\in\Sigma_A)p(\boldsymbol{\theta}) \geq 0$.

Then (i)⇒(ii)⇒(iii)⇒(iv).

Proof of Proposition 28. The first implication is a direct consequence of Proposition 27: we infer from (i) and the continuity of p that $\min p > 0$ and therefore, by Proposition 27, that $\lim_{n\to+\infty, n\geq\deg(p)} \min b_p^n = \min p > 0$, which implies (ii).

To prove that (ii)⇒(iii), assume that there is some $n \geq \deg(p)$ such that $b_p^n > 0$, and consider any $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_A)$. Then since $\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{m} \in \mathcal{N}_A^n$ [BB3], and since by assumption $b_p^n \geq 0$ and $b_p^n(\boldsymbol{\mu}) > 0$ for some $\boldsymbol{\mu} \in \mathcal{N}_A^n$, we see that

$$p(\boldsymbol{\theta}) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \geq b_p^n(\boldsymbol{\mu})\mathrm{B}_{A,\boldsymbol{\mu}}(\boldsymbol{\theta}) > 0.$$

The third implication is an immediate consequence of the continuity of p. □

The following counterexample shows that not necessarily $\mathcal{V}^{+}(A) = \mathcal{V}^{++}(A)$.

Running Example. We go back to the polynomial q on $\Sigma_{\{H,T,U\}}$ defined in Equation (41):

$$q(\boldsymbol{\theta}) = \theta_H^2 - \theta_H\theta_T + \theta_T^2 = (\theta_H - \theta_T)^2 + \theta_H\theta_T \text{ for all } \boldsymbol{\theta} \in \Sigma_{\{H,T,U\}}.$$

We have already argued that this polynomial is not Bernstein positive. Nevertheless, it is obviously positive on the interior of $\Sigma_{\{H,T,U\}}$. $\lozenge$

It is also quite easy to trace the effect on the Bernstein expansion of multiplying with a Bernstein basis polynomial:

Proposition 29. *For all polynomials p on Σ_A, all natural $n \geq \deg(p)$, all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ and all $\boldsymbol{n} \in \mathcal{N}_A^{n+m_A}$:*

$$b_{p\mathrm{B}_{A,\boldsymbol{m}}}^{n+m_A}(\boldsymbol{n}) = \begin{cases} b_p^n(\boldsymbol{n}-\boldsymbol{m})\dfrac{\nu(\boldsymbol{n})}{\nu(\boldsymbol{n}-\boldsymbol{m})\nu(\boldsymbol{m})} & \textit{if } \boldsymbol{n} \geq \boldsymbol{m} \\ 0 & \textit{otherwise.} \end{cases}$$

Proof of Proposition 29. Observe that:

$$\begin{aligned} p\mathrm{B}_{A,\boldsymbol{m}} &= \Bigg(\sum_{\boldsymbol{\mu}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{\mu})\mathrm{B}_{A,\boldsymbol{\mu}}\Bigg)\mathrm{B}_{A,\boldsymbol{m}} = \sum_{\boldsymbol{\mu}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{\mu})\mathrm{B}_{A,\boldsymbol{\mu}}\mathrm{B}_{A,\boldsymbol{m}} \\ &= \sum_{\boldsymbol{\mu}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{\mu})\frac{\nu(\boldsymbol{\mu}+\boldsymbol{m})}{\nu(\boldsymbol{\mu})\nu(\boldsymbol{m})}\mathrm{B}_{A,\boldsymbol{\mu}+\boldsymbol{m}}, \end{aligned}$$

and use the uniqueness of the (Bernstein) basis expansion. $\square$

This allows us to prove the following simple but interesting result about Bernstein positivity:

Proposition 30. *Consider any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ and any polynomial p on Σ_A. Then:*

$$p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^{+}(A) \Leftrightarrow p \in \mathcal{V}^{+}(A).$$

Proof of Proposition 30. First, assume that $p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^{+}(A)$, so there is some natural $n \geq \deg(p)$ such that $b_{p\mathrm{B}_{A,\boldsymbol{m}}}^{n+m_A} > 0$. Then it follows from Proposition 29 that also $b_p^n > 0$, and therefore $p \in \mathcal{V}^{+}(A)$.

Assume, conversely, that $p \in \mathcal{V}^{+}(A)$, so there is some $n \geq \deg(p)$ such that $b_p^n > 0$. Then it follows from Proposition 29 that also $b_{p\mathrm{B}_{A,\boldsymbol{m}}}^{n+m_A} > 0$, and therefore $p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^{+}(A)$. $\square$

Appendix C. The Dirichlet Distribution

The density $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ of the *Dirichlet distribution* with hyperparameter $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A$ is given by:

$$\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{\alpha}) := \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)}\prod_{x\in A}\theta_x^{\alpha_x-1} \text{ for all } \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A),$$

and for any polynomial p on Σ_A we define the corresponding expectation as:[43]

$$\mathrm{Di}_A(p|\boldsymbol{\alpha}) := \int_{\Sigma_A} p(\boldsymbol{\theta}) \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)} \prod_{x\in A} \theta_x^{\alpha_x - 1}\,\mathrm{d}\boldsymbol{\theta}.$$

In particular,

$$\begin{aligned}\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha}) &= \int_{\Sigma_A} \binom{n}{\boldsymbol{m}} \prod_{x\in A} \theta_x^{m_x} \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)} \prod_{x\in A} \theta_x^{\alpha_x - 1}\,\mathrm{d}\boldsymbol{\theta} \\ &= \binom{n}{\boldsymbol{m}} \frac{\Gamma(\alpha_A)}{\Gamma(n+\alpha_A)} \prod_{x\in A} \frac{\Gamma(m_x+\alpha_x)}{\Gamma(\alpha_x)} = \frac{1}{\alpha_A{}^{(n)}} \binom{n}{\boldsymbol{m}} \prod_{x\in A} \alpha_x{}^{(m_x)}, \qquad (80)\end{aligned}$$

using the *ascending factorial* $\alpha^{(r)} := \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)} = \alpha(\alpha+1)\dots(\alpha+r-1)$, with $\alpha \in \mathbb{R}$ and $r \in \mathbb{N}_0$.

The Dirichlet distribution can be used as a prior in combination with a multinomial likelihood, leading to the so-called *Dirichlet multinomial distribution*, which can be described as follows. The probability of observing (a sample with $n \geq 0$ observations with) count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ in a multinomial process with Dirichlet prior density $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ is given by:

$$\begin{aligned}\mathrm{DiMn}_A^n(\{\boldsymbol{m}\}|\boldsymbol{\alpha}) &:= \int_{\Sigma_A} \mathrm{CoMn}_A^n(\{\boldsymbol{m}\}|\boldsymbol{\theta})\,\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Sigma_A} \mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta})\,\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\theta} = \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha}),\end{aligned}$$

where the second equality follows from Equation (16). Therefore, more generally, if we take the expansion of the polynomial p in Bernstein basis polynomials of degree $n \geq \deg(p)$:

$$\begin{aligned}\mathrm{Di}_A(p|\boldsymbol{\alpha}) &= \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\,\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha}) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\,\mathrm{DiMn}_A^n(\mathbb{I}_{\{\boldsymbol{m}\}}|\boldsymbol{\alpha}) \\ &= \mathrm{DiMn}_A^n\left(\sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\mathbb{I}_{\{\boldsymbol{m}\}}\middle|\boldsymbol{\alpha}\right) = \mathrm{DiMn}_A^n(b_p^n|\boldsymbol{\alpha}),\end{aligned}$$

which is the Dirichlet multinomial expectation of the count gamble b_p^n. This is the general and useful relationship between the Dirichlet expectation of a polynomial p, and the Dirichlet multinomial expectation of its Bernstein expansion b_p^n. Although these expectations are strictly speaking only defined for $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A$, we can extend their definition continuously to elements $\boldsymbol{\alpha}$ of $\mathbb{R}_{\geq}^A \setminus \{0\}$ by taking appropriate limits, as Equation (80) indicates.

C.1 Special Properties of the Dirichlet Distribution

We now recall a few interesting properties of the Dirichlet distribution. We begin with the updating property:

Proposition 31 (Updating)**.** *For any category set A, any polynomial $p \in \mathcal{V}(A)$, any count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A$:*

$$\mathrm{Di}_A(p\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha}) = \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha})\,\mathrm{Di}_A(p|\boldsymbol{m}+\boldsymbol{\alpha}).$$

43. The integrals in this section can be interpreted as multiple Riemann integrals.

Proof of Proposition 31.

$$\begin{aligned}
\mathrm{Di}_A(p\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha}) &= \int_{\Sigma_A} p(\boldsymbol{\theta})\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta})\,\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\theta}\\
&= \int_{\Sigma_A} p(\boldsymbol{\theta})\binom{m_A}{\boldsymbol{m}}\prod_{x\in A}\theta_x^{m_x}\frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)}\prod_{x\in A}\theta_x^{\alpha_x-1}\,\mathrm{d}\boldsymbol{\theta}\\
&= \binom{m_A}{\boldsymbol{m}}\frac{\Gamma(\alpha_A)}{\Gamma(m_A+\alpha_A)}\prod_{x\in A}\frac{\Gamma(m_x+\alpha_x)}{\Gamma(\alpha_x)}\int_{\Sigma_A} p(\boldsymbol{\theta})\,\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{m}+\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\theta}\\
&= \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|\boldsymbol{\alpha})\,\mathrm{Di}_A(p|\boldsymbol{m}+\boldsymbol{\alpha}),
\end{aligned}$$

where the last equality follows from Equation (80). □

Next, we turn to the so-called renaming property:

Proposition 32 (Renaming)**.** *For any category sets A and C such that there is some bijective (one-to-one and onto) map $\lambda\colon A\to C$, any polynomial $p\in\mathcal{V}(C)$ and any $\boldsymbol{\alpha}\in\mathbb{R}^A_{>0}$:*

$$\mathrm{Di}_A(p\circ R_\lambda|\boldsymbol{\alpha}) = \mathrm{Di}_C(p|R_\lambda(\boldsymbol{\alpha})).$$

Proof of Proposition 32. Due to the linear nature of the Dirichlet expectation, it clearly suffices to prove the property for the Bernstein basis polynomials $p=\mathrm{B}_{C,\boldsymbol{m}}$, where $\boldsymbol{m}\in\mathcal{N}_C\cup\{\boldsymbol{0}\}$. Observe that R_λ is a bijection too. Then, using Equation (80), if we let $\boldsymbol{\beta}\coloneqq R_\lambda(\boldsymbol{\alpha})$ and $\boldsymbol{n}\coloneqq R_\lambda^{-1}(\boldsymbol{m})$, so $\beta_z=\alpha_{\lambda^{-1}(z)}$ and $m_z=n_{\lambda^{-1}(z)}$ for all $z\in C$, and $\alpha_A=\beta_C$ and $n_A=m_C$, we get:

$$\begin{aligned}
\mathrm{Di}_C(\mathrm{B}_{C,\boldsymbol{m}}|R_\lambda(\boldsymbol{\alpha})) = \mathrm{Di}_C(\mathrm{B}_{C,\boldsymbol{m}}|\boldsymbol{\beta}) &= \binom{m_C}{\boldsymbol{m}}\frac{\Gamma(\beta_C)}{\Gamma(m_C+\beta_C)}\prod_{z\in C}\frac{\Gamma(m_z+\beta_z)}{\Gamma(\beta_z)}\\
&= \binom{n_A}{\boldsymbol{n}}\frac{\Gamma(\alpha_A)}{\Gamma(n_A+\alpha_A)}\prod_{z\in C}\frac{\Gamma(n_{\lambda^{-1}(z)}+\alpha_{\lambda^{-1}(z)})}{\Gamma(\alpha_{\lambda^{-1}(z)})}\\
&= \binom{n_A}{\boldsymbol{n}}\frac{\Gamma(\alpha_A)}{\Gamma(n_A+\alpha_A)}\prod_{x\in A}\frac{\Gamma(n_x+\alpha_x)}{\Gamma(\alpha_x)} = \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{n}}|\boldsymbol{\alpha}),
\end{aligned}$$

and if we take into account that for all $\boldsymbol{\theta}\in\mathrm{int}(\Sigma_A)$:

$$\begin{aligned}
(\mathrm{B}_{C,\boldsymbol{m}}\circ R_\lambda)(\boldsymbol{\theta}) &= \mathrm{B}_{C,\boldsymbol{m}}(R_\lambda(\boldsymbol{\theta}))\\
&= \binom{m_C}{\boldsymbol{m}}\prod_{z\in C}R_\lambda(\theta)_z^{m_z} = \binom{m_C}{\boldsymbol{m}}\prod_{z\in C}\theta_{\lambda^{-1}(z)}^{m_z} = \binom{m_C}{\boldsymbol{m}}\prod_{z\in C}\theta_{\lambda^{-1}(z)}^{n_{\lambda^{-1}(z)}}\\
&= \binom{n_A}{\boldsymbol{n}}\prod_{x\in A}\theta_x^{n_x} = \mathrm{B}_{A,\boldsymbol{n}}(\boldsymbol{\theta}),
\end{aligned}$$

we see that indeed $\mathrm{Di}_C(\mathrm{B}_{C,\boldsymbol{m}}|R_\lambda(\boldsymbol{\alpha})) = \mathrm{Di}_A(\mathrm{B}_{C,\boldsymbol{m}}\circ R_\lambda|\boldsymbol{\alpha})$. □

The so-called pooling property generalises the renaming property:

Proposition 33 (Pooling)**.** *For any category sets A and D such that there is some onto map $\rho\colon A \to D$, any polynomial $p \in \mathcal{V}(D)$ and any $\boldsymbol{\alpha} \in \mathbb{R}^A_{>0}$:*

$$\mathrm{Di}_A(p \circ R_\rho|\boldsymbol{\alpha}) = \mathrm{Di}_D(p|R_\rho(\boldsymbol{\alpha})).$$

Proof of Proposition 33. Due to the linear nature of the Dirichlet expectation, it again suffices to prove the property for the Bernstein basis polynomials $p = \mathrm{B}_{D,\boldsymbol{m}}$, where $\boldsymbol{m} \in \mathcal{N}_D \cup \{\mathbf{0}\}$. Also, if we take into account the renaming property of Proposition 32, it is enough to consider the following special case, where we have some non-empty set D_o and different categories b, c and d not belonging to it, let $A \coloneqq D_o \cup \{b, c\}$ and $D \coloneqq D_o \cup \{d\}$, and define ρ by letting $\rho(x) \coloneqq x$ if $x \in D_o$ and $\rho(b) = \rho(c) \coloneqq d$.

Then on the one hand, taking into account Equation (80), letting $\boldsymbol{\beta} \coloneqq R_\rho(\boldsymbol{\alpha})$:

$$\begin{aligned}\mathrm{Di}_D(\mathrm{B}_{D,\boldsymbol{m}}|R_\rho(\boldsymbol{\alpha})) = \mathrm{Di}_D(\mathrm{B}_{D,\boldsymbol{m}}|\boldsymbol{\beta}) &= \binom{m_D}{\boldsymbol{m}} \frac{\Gamma(\beta_D)}{\Gamma(m_D + \beta_D)} \prod_{z\in D} \frac{\Gamma(m_z + \beta_z)}{\Gamma(\beta_z)} \\ &= \binom{m_D}{\boldsymbol{m}} \frac{\Gamma(\beta_D)}{\Gamma(m_D + \beta_D)} \frac{\Gamma(m_d + \beta_d)}{\Gamma(\beta_d)} \prod_{z\in D_o} \frac{\Gamma(m_z + \beta_z)}{\Gamma(\beta_z)}.\end{aligned} \tag{81}$$

On the other hand,

$$\begin{aligned}&\mathrm{Di}_A(\mathrm{B}_{D,\boldsymbol{m}} \circ R_\rho|\boldsymbol{\alpha}) \\ &\quad= \int_{\Sigma_A} \binom{m_D}{\boldsymbol{m}} (\theta_b + \theta_c)^{m_d} \Bigg(\prod_{z\in D_o} \theta_z^{m_z}\Bigg) \,\mathrm{di}_A(\boldsymbol{\theta}|\boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{\theta} \\ &\quad= \binom{m_D}{\boldsymbol{m}} \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)} \int_{\Sigma_A} (\theta_b + \theta_c)^{m_d} \theta_b^{\alpha_b - 1} \theta_c^{\alpha_c - 1} \prod_{z\in D_o} \theta_z^{m_z + \alpha_z - 1} \,\mathrm{d}\boldsymbol{\theta} \\ &\quad= \binom{m_D}{\boldsymbol{m}} \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)} \sum_{k=0}^{m_d} \binom{m_d}{k} \int_{\Sigma_A} \theta_b^{k+\alpha_b-1} \theta_c^{m_d - k + \alpha_c - 1} \prod_{z\in D_o} \theta_z^{m_z + \alpha_z - 1} \,\mathrm{d}\boldsymbol{\theta} \\ &\quad= \binom{m_D}{\boldsymbol{m}} \frac{\Gamma(\alpha_A)}{\prod_{x\in A}\Gamma(\alpha_x)} \sum_{k=0}^{m_d} \binom{m_d}{k} \frac{\Gamma(k + \alpha_b)\Gamma(m_d - k + \alpha_c) \prod_{z\in D_o}\Gamma(m_z + \alpha_z)}{\Gamma(m_D + \alpha_A)}.\end{aligned}$$

So, if we compare both results and recall that $\beta_D = \alpha_A$, $\alpha_z = \beta_z$ for $z \in D_o$ and $\beta_d = \alpha_b + \alpha_c$, we see that we must prove that:

$$\frac{\Gamma(m_d + \alpha_b + \alpha_c)}{\Gamma(\alpha_b + \alpha_c)} = \frac{1}{\Gamma(\alpha_b)\Gamma(\alpha_c)} \sum_{k=0}^{m_d} \binom{m_d}{k} \Gamma(k + \alpha_b)\Gamma(m_d - k + \alpha_c)$$

or equivalently, using ascending factorials:

$$(\alpha_b + \alpha_c)^{(m_d)} = \sum_{k=0}^{m_d} \binom{m_d}{k} {\alpha_b}^{(k)} {\alpha_c}^{(m_d - k)}. \tag{82}$$

So we see that proving the pooling property is essentially equivalent to proving Equation (82), which is the ‘binomial theorem for ascending factorials’. This is a well-known result, and it follows from the fact that ascending factorials are Sheffer sequences of binomial type (Sheffer, 1939). For completeness, we give a proof for it here, which is now very easy, because we have just shown that it will hold if we can prove the pooling property in the particular case that $D_o = \{a\}$, where a is a category different from b, c and d. So $A = \{a, b, c\}$ and $D = \{a, d\}$, and in this case we can rewrite Equation (81) as:

$$\begin{aligned}
&\mathrm{Di}_D(\mathrm{B}_{D,\boldsymbol{m}}|R_\rho(\boldsymbol{\alpha})) \\
&\qquad = \binom{m_a + m_d}{m_a} \frac{\Gamma(\alpha_a + \alpha_b + \alpha_c)}{\Gamma(m_a + m_d + \alpha_a + \alpha_b + \alpha_c)} \frac{\Gamma(m_d + \alpha_b + \alpha_c)}{\Gamma(\alpha_b + \alpha_c)} \frac{\Gamma(m_a + \alpha_a)}{\Gamma(\alpha_a)}
\end{aligned}$$

whereas

$$\mathrm{Di}_A(\mathrm{B}_{D,\boldsymbol{m}} \circ R_\rho|\boldsymbol{\alpha}) = \binom{m_a + m_d}{m_a} \frac{\Gamma(\alpha_a + \alpha_b + \alpha_c)}{\Gamma(\alpha_a)\Gamma(\alpha_b)\Gamma(\alpha_c)} I$$

where we let

$$\begin{aligned}
I &\coloneqq \int_0^1 \left(\int_0^{1-\theta_a} (1-\theta_a)^{m_d} \theta_b^{\alpha_b - 1} (1 - \theta_a - \theta_b)^{\alpha_c - 1} \theta_a^{m_a + \alpha_a - 1} \,\mathrm{d}\theta_b \right) \mathrm{d}\theta_a \\
&= \int_0^1 (1-\theta_a)^{m_d} \theta_a^{m_a + \alpha_a - 1} \left(\int_0^{1-\theta_a} \theta_b^{\alpha_b - 1} (1 - \theta_a - \theta_b)^{\alpha_c - 1} \,\mathrm{d}\theta_b \right) \mathrm{d}\theta_a \\
&= \int_0^1 (1-\theta_a)^{m_d + \alpha_b + \alpha_c - 1} \theta_a^{m_a + \alpha_a - 1} \left(\int_0^1 t^{\alpha_b - 1} (1-t)^{\alpha_c - 1} \,\mathrm{d}t \right) \mathrm{d}\theta_a \\
&= B(m_a + \alpha_a, m_d + \alpha_b + \alpha_c) B(\alpha_b, \alpha_c) = \frac{\Gamma(m_a + \alpha_a)\Gamma(m_d + \alpha_b + \alpha_c)}{\Gamma(m_a + m_d + \alpha_a + \alpha_b + \alpha_c)} \frac{\Gamma(\alpha_b)\Gamma(\alpha_c)}{\Gamma(\alpha_b + \alpha_c)},
\end{aligned}$$

using the well-known evaluation of the Beta function in terms of Gamma functions. □

Finally, we look at properties related to restriction.

Proposition 34 (Restriction)**.** *For any category sets A and B such that $B \subseteq A$, any polynomial $p \in \mathcal{V}(B)$, any $\boldsymbol{\alpha} \in \mathbb{R}^A_{>0}$ and any $r \in \mathbb{N}_0$:*

$$\mathrm{Di}_A(\mathrm{I}^r_{B,A}(p)|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_A)}{\Gamma(\deg(p) + r + \alpha_A)} \frac{\Gamma(\deg(p) + r + \alpha_B)}{\Gamma(\alpha_B)} \mathrm{Di}_B(p|r_B(\boldsymbol{\alpha})).$$

Proof of Proposition 34. Let $n \coloneqq \deg(p) + r$, then due to the linearity of the Dirichlet expectation operator, and Equations (29) and (80):

$$\begin{aligned}
\mathrm{Di}_A(\mathrm{I}^r_{B,A}(p)|\boldsymbol{\alpha}) &= \sum_{\boldsymbol{n} \in \mathcal{N}^n_B} b^n_p(\boldsymbol{n}) \,\mathrm{Di}_A(\mathrm{B}_{A,i_A(\boldsymbol{n})}|\boldsymbol{\alpha}) \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}^n_B} b^n_p(\boldsymbol{n}) \binom{n}{\boldsymbol{n}} \frac{\Gamma(\alpha_A)}{\Gamma(n + \alpha_A)} \frac{\prod_{x \in B} \Gamma(n_x + \alpha_x) \prod_{x \in A \setminus B} \Gamma(\alpha_x)}{\prod_{x \in B} \Gamma(\alpha_x) \prod_{x \in A \setminus B} \Gamma(\alpha_x)} \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}^n_B} b^n_p(\boldsymbol{n}) \binom{n}{\boldsymbol{n}} \frac{\Gamma(\alpha_A)}{\Gamma(n + \alpha_A)} \prod_{x \in B} \frac{\Gamma(n_x + \alpha_x)}{\Gamma(\alpha_x)} \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}^n_B} b^n_p(\boldsymbol{n}) \frac{\Gamma(\alpha_A)}{\Gamma(n + \alpha_A)} \frac{\Gamma(n + \alpha_B)}{\Gamma(\alpha_B)} \,\mathrm{Di}_B(\mathrm{B}_{B,\boldsymbol{n}}|r_B(\boldsymbol{\alpha}))
\end{aligned}$$

$$= \frac{\Gamma(\alpha_A)}{\Gamma(n+\alpha_A)} \frac{\Gamma(n+\alpha_B)}{\Gamma(\alpha_B)} \operatorname{Di}_B(p|r_B(\boldsymbol{\alpha})),$$

concluding the proof. □

Appendix D. The Original IDMM Inference System by Walley and Bernard

The IDMM inference system Φ^s_{IDM}, as we introduced it in Section 12, differs from the one originally proposed by Walley and Bernard (1999).[44] In this appendix, we discuss the original IDMM inference system, which we denote by Φ^s_{OI}, explain how it is related to ours, and illustrate some of the advantages our version has over the one by Walley and Bernard.

D.1 Defining the Original IDMM Inference System

For any $s \in \mathbb{R}_{>0}$, and any category set A, consider the following set of polynomials:

$$\begin{aligned}\mathcal{H}^s_{\mathrm{OI},A} &\coloneqq \{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{\alpha} \in \mathrm{K}^s_A) \operatorname{Di}_A(p|\boldsymbol{\alpha}) > 0\} \\ &= \{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A)) \operatorname{Di}_A(p|s\boldsymbol{t}) > 0\}.\end{aligned}$$

For reasons that should become clear shortly, we call the inference system Φ^s_{OI} defined by

$$\Phi^s_{\mathrm{OI}}(A) \coloneqq \mathcal{H}^s_{\mathrm{OI},A} \text{ for all category sets } A,$$

the *original IDMM inference system* with hyperparameter $s > 0$. Updating is done in much the same way as for the inference system Φ^s_{IDM} in Section 12. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{H}^s_{\mathrm{OI},A}\rfloor\check{\boldsymbol{m}} = \{p \in \mathcal{V}(A) \colon (\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A)) \operatorname{Di}_A(p|\check{\boldsymbol{m}} + s\boldsymbol{t}) > 0\},$$

and this should be compared with Equation (54). We leave it as an exercise to the reader to check that Φ^s_{OI} is coherent and representation insensitive.[45] However, as illustrated by the counterexample in Section D.3, Φ^s_{OI} is *not specific*.

The predictive models of Φ^s_{OI} are easily derived by mimicking the approach used in Section 12 to derive the predictive models of Φ^s_{IDM}; see Equations (56) and (57). For any $\check{n} \in \mathbb{N}_0$, $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}^{\check{n}}_A$:

$$\mathcal{D}^{s,\hat{n}}_{\mathrm{OI},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A^{\hat{n}}) \colon (\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A)) \operatorname{Di}_A(\operatorname{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + s\boldsymbol{t}) > 0\right\}, \tag{83}$$

and

$$\underline{P}^{s,\hat{n}}_{\mathrm{OI},A}(f|\check{\boldsymbol{m}}) = \inf_{\boldsymbol{t}\in\operatorname{int}(\Sigma_A)} \operatorname{Di}_A(\operatorname{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + s\boldsymbol{t}) \text{ for all gambles } f \text{ on } A^{\hat{n}}. \tag{84}$$

The latter expression motivates why we refer to Φ^s_{OI} as the original IDMM inference system: its predictive lower previsions coincide with those proposed by Walley and Bernard (1999).

44. Strictly speaking, Walley and Bernard did not propose an inference system in our sense, but rather a collection of prior and posterior predictive lower previsions for each category set A. The inference system we call ‘the original IDMM inference system’ is one that produces these predictive lower previsions.

45. The proof is very similar to the one for Φ^s_{IDM} [see Theorem 21].

Using Equation (83) for $\hat{n}=1$, and mimicking the argument in the proof of Equation (59) in Appendix E.7, we see that

$$\mathcal{D}^{s,1}_{\mathrm{OI},A}\rfloor\check{\boldsymbol{m}}=\left\{f\in\mathcal{L}(A)\colon f>-\frac{1}{s}\sum_{x\in A}f(x)\check{m}_x\right\}=\mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}\text{ for all }\check{\boldsymbol{m}}\in\mathcal{N}_A\cup\{\mathbf{0}\}.$$

This tells us that the IDMM and the original IDMM have the same *immediate* prediction models. The corresponding immediate predictive lower previsions for the original IDMM are well-known and are of course identical to the ones produced by our version of the IDMM inference system, as given by Equation (60). However, as the examples in the next section illustrate, this equality does not extend beyond immediate prediction: the IDMM and the original IDMM are different coherent inference systems, which leads us to the general and important conclusion that *coherent inference systems are not completely determined by their immediate prediction models.*

Nevertheless, both approaches are closely related; by comparing Equations (84) and (57), we see that for any $\check{n}\in\mathbb{N}_0$, $\hat{n}\in\mathbb{N}$ and any $\check{\boldsymbol{m}}\in\mathcal{N}^{\check{n}}_A$

$$\underline{P}^{s,\hat{n}}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}})=\inf_{0<s'<s}\underline{P}^{s',\hat{n}}_{\mathrm{OI},A}(f|\check{\boldsymbol{m}})\text{ for all gambles }f\text{ on }A^{\hat{n}}.\tag{85}$$

D.2 The Original IDMM Inference System Is Not Monotone in s

The hyperparameter s of the original IDMM inference system is usually interpreted as a degree of caution. Higher values of s are often claimed to produce inferences that are more cautious and less informative. The following quote from Walley and Bernard (1999, Section 2.4) makes this explicit:

> "If B is any event concerning future observations, the IDMM(s) produces intervals of posterior probabilities $[\underline{P}(B|\boldsymbol{x}),\overline{P}(B|\boldsymbol{x})]$ which are nested and become wider as s increases. This means that the inferences produced by two IDMMs with different values of s are always consistent with each other, and the effect of increasing s is simply to make inferences more cautious and less informative."

Similar statements can be found in related papers by Walley (1996, Section 2.5) and Bernard (2005, Section 4.6). Although this is indeed true for many inferences, including many important ones—for example, the immediate predictions—, it does not hold for "any event concerning future observations", as illustrated by the following example, where the lower probability of an event concerning two future observations is shown to initially increase with s.

Example 1. Consider a situation where the possibility space A consists of two elements only, say heads (H) and tails (T), each of which has been observed once, so $\check{n}=2$ and $\check{\boldsymbol{m}}=(\check{m}_H,\check{m}_T)=(1,1)$. We are interested in the predictive lower probability that during the next two trials, heads and tails will each be observed once: so $\hat{n}=2$ and we are looking for the predictive lower probability of the event $[\hat{\boldsymbol{m}}]=\{(H,T),(T,H)\}$, with $\hat{\boldsymbol{m}}=(\hat{m}_H,\hat{m}_T)=(1,1)$. For the original IDMM inference system, the following formula provides a closed-form expression:

$$\underline{P}^{s,\hat{n}}_{\mathrm{OI},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}})=\inf_{\boldsymbol{t}\in\mathrm{int}(\Sigma_A)}\mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(\mathbb{I}_{[\hat{\boldsymbol{m}}]})|\check{\boldsymbol{m}}+s\boldsymbol{t})=\inf_{\boldsymbol{t}\in\mathrm{int}(\Sigma_A)}\mathrm{Di}_A(\mathrm{B}_{A,\hat{\boldsymbol{m}}}|\check{\boldsymbol{m}}+s\boldsymbol{t})$$

$$= \inf_{\boldsymbol{t}\in\operatorname{int}(\Sigma_A)} \frac{1}{(\check{n}+s)^{(\hat{n})}} \binom{\hat{n}}{\hat{\boldsymbol{m}}} \prod_{x\in A} (\check{m}_x + st_x)^{(\hat{m}_x)}$$

$$= \inf_{0<t<1} \frac{2(1+st)(1+s(1-t))}{(2+s)(3+s)} = \frac{2(1+s)}{(2+s)(3+s)}. \tag{86}$$

We conclude that $\underline{P}^{s,\hat{n}}_{\mathrm{OI},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}})$ initially increases with s; see also Figure 2. ◊

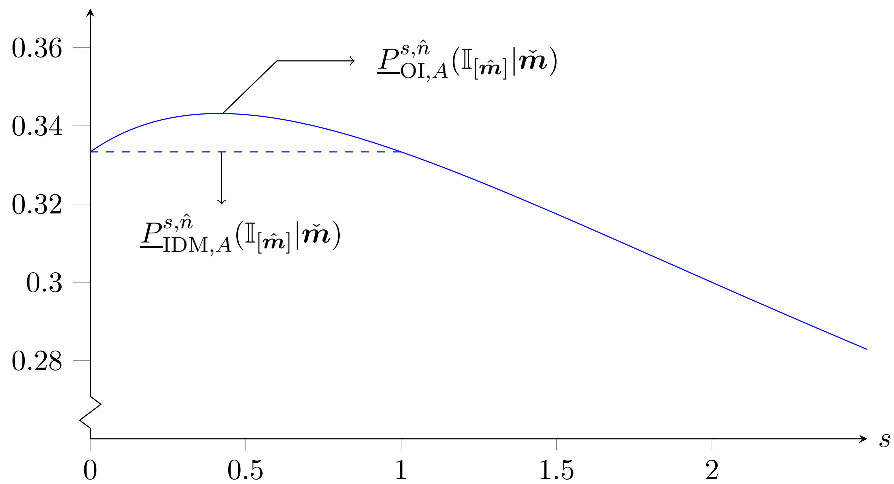

Figure 2: Lower probability of observing two different outcomes during the next two experiments, given that the possibility space consists of two categories, each of which has already been observed once: solutions according to Φ^s_{OI} (solid line) and Φ^s_{IDM} (dashed and solid line); see Examples 1 and 2 for more information.

For our version of the IDMM inference system, the statement made in the aforementioned quote does hold for "any event concerning future observations". This follows trivially from Equation (85). We illustrate this in our next example.

Example 2. Consider again the problem in Example 1. This time, we solve it using our version of the IDMM. The result is also depicted in Figure 2, as a function of the hyperparameter s. In contrast with $\underline{P}^{s,\hat{n}}_{\mathrm{OI},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}})$, $\underline{P}^{s,\hat{n}}_{\mathrm{IDM},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}})$ is a non-increasing function of s. Indeed,

$$\underline{P}^{s,\hat{n}}_{\mathrm{IDM},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}}) = \begin{cases} \lim_{s'\to 0} \underline{P}^{s',\hat{n}}_{\mathrm{OI},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}}) = \dfrac{1}{3} & \text{if } 0 < s < 1 \\ \underline{P}^{s,\hat{n}}_{\mathrm{OI},A}(\mathbb{I}_{[\hat{\boldsymbol{m}}]}|\check{\boldsymbol{m}}) = \dfrac{2(1+s)}{(2+s)(3+s)} & \text{if } s \geq 1 \end{cases}$$

is the closed-form expression we find by combining Equations (85) and (86). ◊

Clearly, the inferences for Φ^s_{OI} and Φ^s_{IDM} can differ: it suffices to compare the results in Examples 1 and 2; see Figure 2 as well. Therefore, it seems clear that Walley's (1996, p. 51) statement that "[...] s can be allowed to vary between 0 and $\overline{s}$, and this produces exactly the same inferences as the IDM with $s = \overline{s}$." or equivalently, that Φ^s_{OI} and Φ^s_{IDM} produce the same inferences, should be taken to apply to immediate prediction only.

D.3 The Original IDMM Inference System Is Not Specific

As announced in Theorem 21, our version of the IDMM inference system is specific. We now show that, at least for some values of the hyperparameter s, this is not true for the original version.

Consider any $\check{n} \in \mathbb{N}_0$, $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_B^{\check{n}}$. Then for all $A \supseteq B$ and all $f \in \mathcal{L}(B^{\hat{n}})$:

$$\begin{aligned} f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_{\mathrm{OI},A}^{s,\hat{n}} \rfloor i_A(\check{\boldsymbol{m}}) &\Leftrightarrow (\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A))\, \mathrm{Di}_A(\mathrm{Mn}_A^{\hat{n}}(f\mathbb{I}_{B^{\hat{n}}})|i_A(\check{\boldsymbol{m}}) + s\boldsymbol{t}) > 0 \\ &\Leftrightarrow (\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A))\, \mathrm{Di}_B(\mathrm{Mn}_B^{\hat{n}}(f)|\check{\boldsymbol{m}} + sr_B(\boldsymbol{t})) > 0, \end{aligned}$$

where the last equivalence is a consequence of Propositions 41 and 34 and the fact that $r_B(i_A(\check{\boldsymbol{m}})) = \check{\boldsymbol{m}}$. If in particular $B \subset A$, it is not hard to see that $\Delta_B^s = \{sr_B(\boldsymbol{t}) \colon \boldsymbol{t} \in \operatorname{int}(\Sigma_A)\}$, which implies that:

$$f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_{\mathrm{OI},A}^{s,\hat{n}} \rfloor i_A(\check{\boldsymbol{m}}) \Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_B^s)\, \mathrm{Di}_B(\mathrm{Mn}_B^{\hat{n}}(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0 \Leftrightarrow f \in \mathcal{D}_{\mathrm{IDM},B}^{s,\hat{n}} \rfloor \check{\boldsymbol{m}},$$

and therefore also:

$$\underline{P}_{\mathrm{OI},A}^{s,\hat{n}}(f|i_A(\check{\boldsymbol{m}}), B^{\hat{n}}) = \inf_{\boldsymbol{\alpha} \in \Delta_B^s} \mathrm{Di}_B(\mathrm{Mn}_B^{\hat{n}}(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) = \underline{P}_{\mathrm{IDM},B}^{s,\hat{n}}(f|\check{\boldsymbol{m}}).$$

On the other hand, due to Equation (SP1), if Φ_{OI}^s were specific, we would have that:

$$\underline{P}_{\mathrm{OI},A}^{s,\hat{n}}(f|i_A(\check{\boldsymbol{m}}), B^{\hat{n}}) = \underline{P}_{\mathrm{OI},B}^{s,\hat{n}}(f|r_B(i_A(\check{\boldsymbol{m}}))) = \underline{P}_{\mathrm{OI},B}^{s,\hat{n}}(f|\check{\boldsymbol{m}}).$$

Hence, in order for Φ_{OI}^s to be specific, it is necessary for $\underline{P}_{\mathrm{OI},B}^{s,\hat{n}}(\cdot|\check{\boldsymbol{m}})$ and $\underline{P}_{\mathrm{IDM},B}^{s,\hat{n}}(\cdot|\check{\boldsymbol{m}})$ to coincide. As illustrated by the examples in the previous section, this is not necessarily the case. Therefore, Φ_{OI}^s is not always specific. In the counterexample we have provided, the difference occurs for $s < 1$ only, whereas in practice, s is usually chosen to be either 1 or 2 (Walley & Bernard, 1999, Section 2.4). It would be interesting to see whether similar counterexamples can be constructed for $s \geq 1$.

That the original IDMM inference systems are not specific, apparently contradicts Theorem 11 by De Cooman et al. (2009a), which seems to state that they are. But in fact, what that theorem states is that the original IDMM *immediate prediction* models satisfy a weaker specificity condition, tailored to immediate prediction only. Since the immediate prediction models for the original IDMM and the IDMM coincide, there is no contradiction.

Appendix E. Proofs and Additional Results That Are More Technical

E.1 Proofs of Results in Section 4

Proof of Theorem 4. For the sake of notational simplicity, we use the intuitive notation $f(X_k)$ for $\mathrm{ext}_k^{\hat{n}}(f)$. We give the proof for the most general definition, in terms of sets of desirable gambles. The proof for lower previsions then follows immediately.

Consider any category set A, any $\hat{n} \in \mathbb{N}$, any $1 \leq k \leq \hat{n}$ and any gamble f on A such that $f(X_k) \in \mathcal{D}_A^{\hat{n}}$—we may assume without loss of generality that A is not a singleton. This already implies that $f \not\leq 0$, by coherence [D4]. Hence in particular $f \neq 0$ and $\max f > 0$. Assume *ex absurdo* that $f \not\geq 0$, then there must be some $a \in A$ for which $f(a) < 0$. Define the gamble g on A by letting $g(a) \coloneqq f(a)$ and $g(x) \coloneqq \max f > 0$ for all $x \in A \setminus \{a\}$. Then $g \geq f$ and therefore $g(X_k) \geq f(X_k)$, which implies, by coherence [use D2 and D3], that also $g(X_k) \in \mathcal{D}_A^{\hat{n}}$. If we now let $\lambda \coloneqq \max f - f(a) > 0$ and $\delta \coloneqq -f(a)/\lambda > 0$, and define the gamble $h \coloneqq g/\lambda = -\delta + \mathbb{I}_{A\setminus\{a\}}$, then also, by coherence [D3], $h(X_k) \in \mathcal{D}_A^{\hat{n}}$, because $\lambda > 0$.

Now consider any natural number $N \geq 2$, then it follows from repeatedly applying pooling and renaming invariance in an appropriate manner that $-\delta + \mathbb{I}_{\{a_1\}}(Z_k) \in \mathcal{D}^{\hat{n}}_{\{a_1,\dots,a_N\}}$, where Z_k is any variable that assumes the value a_1 when $X_k \neq a$ and that assumes some value in $\{a_2,\dots,a_N\}$ when $X_k = a$. By repeatedly applying category permutation invariance, we find that $-\delta + \mathbb{I}_{\{a_\ell\}}(Z_k) \in \mathcal{D}^{\hat{n}}_{\{a_1,\dots,a_N\}}$ for all $\ell \in \{1,\dots,N\}$. Coherence [D3] then tells us that $-N\delta + 1 = \sum_{\ell=1}^{N}[-\delta + \mathbb{I}_{\{a_\ell\}}(Z_k)] \in \mathcal{D}^{\hat{n}}_{\{a_1,\dots,a_N\}}$. This leads to a contradiction with coherence [D4] if we choose N large enough. □

E.2 Proofs of Results in Section 7

Proposition 35. *For all $n \in \mathbb{N}$ and $\boldsymbol{x} \in A^n$: $\boldsymbol{T}(\rho\boldsymbol{x}) = R_\rho(\boldsymbol{T}(\boldsymbol{x}))$.*

Proof of Proposition 35. Consider any $z \in D$, then

$$T_z(\rho\boldsymbol{x}) = |\{k \in \{1,\dots,n\}\colon \rho(x_k) = z\}| = \sum_{y\in A\colon \rho(y)=z} |\{k \in \{1,\dots,n\}\colon x_k = y\}| = \sum_{y\in A\colon \rho(y)=z} T_y(\boldsymbol{x}) = R_\rho(\boldsymbol{T}(\boldsymbol{x}))_z,$$

concluding the proof. □

Lemma 36. *For all $n \in \mathbb{N}$, all $\boldsymbol{m} \in \mathcal{N}_A^n$ and all $\boldsymbol{y} \in D^n$:*

$$\frac{1}{\nu(\boldsymbol{m})}\sum_{\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho\boldsymbol{x}\}}(\boldsymbol{y}) = \frac{1}{\nu(R_\rho(\boldsymbol{m}))}\mathbb{I}_{[R_\rho(\boldsymbol{m})]}(\boldsymbol{y}).$$

Proof of Lemma 36. Consider the map $M_{\boldsymbol{m}}\colon D^n \to \mathbb{R}$ defined by $M_{\boldsymbol{m}} \coloneqq \sum_{\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho\boldsymbol{x}\}}$. Then for any permutation π of the index set $\{1,\dots,n\}$ and any $\boldsymbol{y} \in D^n$, we see that

$$M_{\boldsymbol{m}}(\pi\boldsymbol{y}) = \sum_{\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho\boldsymbol{x}\}}(\pi\boldsymbol{y}) = \sum_{\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho(\pi^{-1}\boldsymbol{x})\}}(\boldsymbol{y}) = \sum_{\pi\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho\boldsymbol{x}\}}(\boldsymbol{y}) = \sum_{\boldsymbol{x}\in[\boldsymbol{m}]}\mathbb{I}_{\{\rho\boldsymbol{x}\}}(\boldsymbol{y}) = M_{\boldsymbol{m}}(\boldsymbol{y}),$$

which tells us that $M_{\boldsymbol{m}}$ is permutation invariant and therefore constant on the atoms $[\boldsymbol{n}]$, $\boldsymbol{n} \in \mathcal{N}_D^n$. This means that, with obvious notations, $M_{\boldsymbol{m}} = \sum_{\boldsymbol{n}\in\mathcal{N}_D^n} M_{\boldsymbol{m}}(\boldsymbol{n})\mathbb{I}_{[\boldsymbol{n}]}$. Now $M_{\boldsymbol{m}}(\boldsymbol{y}) > 0$ implies that there is some $\boldsymbol{x} \in [\boldsymbol{m}]$ such that $\boldsymbol{y} = \rho\boldsymbol{x}$, and therefore, by Proposition 35, $\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{T}(\rho\boldsymbol{x}) = R_\rho(\boldsymbol{T}(\boldsymbol{x})) = R_\rho(\boldsymbol{m})$ and therefore $\boldsymbol{y} \in [R_\rho(\boldsymbol{m})]$. This tells us that $M_{\boldsymbol{m}}(\boldsymbol{n}) = 0$ unless $\boldsymbol{n} = R_\rho(\boldsymbol{m})$ and therefore $M_{\boldsymbol{m}} = M_{\boldsymbol{m}}(R_\rho(\boldsymbol{m}))\mathbb{I}_{[R_\rho(\boldsymbol{m})]}$. Now if we plug $f \coloneqq 1$ into Equation (87), we see that

$$\nu(\boldsymbol{m}) = \sum_{\boldsymbol{y}\in D^n} M_{\boldsymbol{m}}(\boldsymbol{y}) = \sum_{\boldsymbol{y}\in D^n} M_{\boldsymbol{m}}(R_\rho(\boldsymbol{m}))\mathbb{I}_{[R_\rho(\boldsymbol{m})]}(\boldsymbol{y}) = M_{\boldsymbol{m}}(R_\rho(\boldsymbol{m}))\nu(R_\rho(\boldsymbol{m})). \qquad \square$$

Lemma 37. *For all $n \in \mathbb{N}$ and all $\boldsymbol{n} \in \mathcal{N}_D^n$:*

$$\mathrm{B}_{D,\boldsymbol{n}} \circ R_\rho = \sum_{\boldsymbol{m} \in \mathcal{N}_A^n \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \mathrm{B}_{A,\boldsymbol{m}}$$

Proof of Lemma 37. For any $\boldsymbol{\theta}$ in Σ_A, we have that

$$\begin{aligned}
(\mathrm{B}_{D,\boldsymbol{n}} \circ R_\rho)(\boldsymbol{\theta}) &= \binom{n}{\boldsymbol{n}} \prod_{z \in D} \Bigg(\sum_{x \in \rho^{-1}(\{z\})} \theta_x \Bigg)^{n_z} \\
&= \binom{n}{\boldsymbol{n}} \prod_{z \in D} \sum_{\boldsymbol{m}^z \in \mathcal{N}^{n_z}_{\rho^{-1}(\{z\})}} \binom{n_z}{\boldsymbol{m}^z} \prod_{x \in \rho^{-1}(\{z\})} \theta_x^{m_x^z} \\
&= \binom{n}{\boldsymbol{n}} \sum_{\boldsymbol{m} \in \mathcal{N}_A^n \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \Bigg(\prod_{x \in A} \theta_x^{m_x} \Bigg) \prod_{z \in D} \binom{n_z}{\boldsymbol{m}_{|\rho^{-1}(\{z\})}} \\
&= \sum_{\boldsymbol{m} \in \mathcal{N}_A^n \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \binom{n}{\boldsymbol{m}} \prod_{x \in A} \theta_x^{m_x} = \sum_{\boldsymbol{m} \in \mathcal{N}_A^n \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}),
\end{aligned}$$

concluding the proof. □

This lemma allows us to prove two related propositions.

Proposition 38. *For all $n \in \mathbb{N}$ and all gambles f on D^n: $\mathrm{Mn}_A^n(f \circ \rho) = \mathrm{Mn}_D^n(f) \circ R_\rho$.*

Proof of Proposition 38. First of all, we have for any count vector $\boldsymbol{m}$ in $\mathcal{N}_A^n$ that

$$\begin{aligned}
\mathrm{Hy}_A^n(f \circ \rho | \boldsymbol{m}) &= \frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x} \in [\boldsymbol{m}]} f(\rho \boldsymbol{x}) = \frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x} \in [\boldsymbol{m}]} \sum_{\boldsymbol{y} \in D^n} \mathbb{I}_{\{\rho \boldsymbol{x}\}}(\boldsymbol{y}) f(\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in D^n} f(\boldsymbol{y}) \frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x} \in [\boldsymbol{m}]} \mathbb{I}_{\{\rho \boldsymbol{x}\}}(\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in D^n} f(\boldsymbol{y}) \frac{1}{\nu(R_\rho(\boldsymbol{m}))} \mathbb{I}_{[R_\rho(\boldsymbol{m})]}(\boldsymbol{y}) \\
&= \frac{1}{\nu(R_\rho(\boldsymbol{m}))} \sum_{\boldsymbol{y} \in [R_\rho(\boldsymbol{m})]} f(\boldsymbol{y}) = \mathrm{Hy}_D^n(f | R_\rho(\boldsymbol{m})),
\end{aligned} \tag{87}$$

where the fourth equality follows from Lemma 36. Therefore indeed:

$$\begin{aligned}
\mathrm{Mn}_A^n(f \circ \rho) &= \sum_{\boldsymbol{m} \in \mathcal{N}_A^n} \mathrm{Hy}_A^n(f \circ \rho | \boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}} = \sum_{\boldsymbol{m} \in \mathcal{N}_A^n} \mathrm{Hy}_D^n(f | R_\rho(\boldsymbol{m})) \mathrm{B}_{A,\boldsymbol{m}} \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}_D^n} \mathrm{Hy}_D^n(f | \boldsymbol{n}) \sum_{\boldsymbol{m} \in \mathcal{N}_A^n \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \mathrm{B}_{A,\boldsymbol{m}} \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}_D^n} \mathrm{Hy}_D^n(f | \boldsymbol{n}) (\mathrm{B}_{D,\boldsymbol{n}} \circ R_\rho) = \mathrm{Mn}_D^n(f) \circ R_\rho,
\end{aligned}$$

where the fourth equality now follows from Lemma 37. □

Proposition 39. *For all polynomials p on Σ_D and all $n \in \mathbb{N}_0$ such that $n \geq \deg(p)$: $b^n_{p \circ R_\rho} = b^n_p \circ R_\rho$.*

Proof of Proposition 39. We find after expanding p in the appropriate Bernstein basis:

$$
\begin{aligned}
p \circ R_\rho &= \Bigg(\sum_{\boldsymbol{n} \in \mathcal{N}^n_D} b^n_p(\boldsymbol{n}) \mathrm{B}_{D,\boldsymbol{n}}\Bigg) \circ R_\rho = \sum_{\boldsymbol{n} \in \mathcal{N}^n_D} b^n_p(\boldsymbol{n})(\mathrm{B}_{D,\boldsymbol{n}} \circ R_\rho) \\
&= \sum_{\boldsymbol{n} \in \mathcal{N}^n_D} b^n_p(\boldsymbol{n}) \sum_{\boldsymbol{m} \in \mathcal{N}^n_A \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} \mathrm{B}_{A,\boldsymbol{m}} = \sum_{\boldsymbol{n} \in \mathcal{N}^n_D} \sum_{\boldsymbol{m} \in \mathcal{N}^n_A \colon R_\rho(\boldsymbol{m}) = \boldsymbol{n}} b^n_p(R_\rho(\boldsymbol{m})) \mathrm{B}_{A,\boldsymbol{m}} \\
&= \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} (b^n_p \circ R_\rho)(\boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}},
\end{aligned}
$$

where the third equality follows from Lemma 37. The desired result now follows from the uniqueness of an expansion in a (Bernstein) basis. ☐

Proof of Theorem 7. Fix any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $\check{n}, \hat{n} \in \mathbb{N}$, any $\check{\boldsymbol{x}} \in A^{\check{n}}$ and any gamble f on $D^{\hat{n}}$. We use the notation $\mathcal{H}_A \coloneqq \Phi(A)$ and $\mathcal{H}_D \coloneqq \Phi(D)$, and we transform Condition (RI2) using the equivalence in Condition (18). On the one hand, letting $\check{\boldsymbol{m}} \coloneqq \boldsymbol{T}(\check{\boldsymbol{x}})$:

$$
\begin{gathered}
f \circ \rho \in \mathcal{D}^{\hat{n}}_A \Leftrightarrow \mathrm{Mn}^{\hat{n}}_A(f \circ \rho) \in \mathcal{H}_A \Leftrightarrow \mathrm{Mn}^{\hat{n}}_D(f) \circ R_\rho \in \mathcal{H}_A \\
f \circ \rho \in \mathcal{D}^{\hat{n}}_A \rfloor \check{\boldsymbol{x}} \Leftrightarrow \mathrm{B}_{A,\check{\boldsymbol{m}}} \mathrm{Mn}^{\hat{n}}_A(f \circ \rho) \in \mathcal{H}_A \Leftrightarrow \mathrm{B}_{A,\check{\boldsymbol{m}}}(\mathrm{Mn}^{\hat{n}}_D(f) \circ R_\rho) \in \mathcal{H}_A,
\end{gathered}
$$

where the second equivalences follow from Proposition 38. On the other hand, recalling that $\boldsymbol{T}(\rho\check{\boldsymbol{x}}) = R_\rho(\boldsymbol{T}(\check{\boldsymbol{x}})) = R_\rho(\check{\boldsymbol{m}})$ by Proposition 35:

$$
\begin{gathered}
f \in \mathcal{D}^{\hat{n}}_D \Leftrightarrow \mathrm{Mn}^{\hat{n}}_D(f) \in \mathcal{H}_D \\
f \in \mathcal{D}^{\hat{n}}_D \rfloor \rho\check{\boldsymbol{x}} \Leftrightarrow \mathrm{B}_{D,R_\rho(\check{\boldsymbol{m}})} \mathrm{Mn}^{\hat{n}}_D(f) \in \mathcal{H}_D.
\end{gathered}
$$

This tells us that the equivalences in Condition (RI2) can be rewritten as:

$$
\begin{gathered}
\mathrm{Mn}^{\hat{n}}_D(f) \circ R_\rho \in \mathcal{H}_A \Leftrightarrow \mathrm{Mn}^{\hat{n}}_D(f) \in \mathcal{H}_D \\
\mathrm{B}_{A,\check{\boldsymbol{m}}}(\mathrm{Mn}^{\hat{n}}_D(f) \circ R_\rho) \in \mathcal{H}_A \Leftrightarrow \mathrm{B}_{D,R_\rho(\check{\boldsymbol{m}})} \mathrm{Mn}^{\hat{n}}_D(f) \in \mathcal{H}_D.
\end{gathered}
$$

The proof is complete if we observe (and recall from the discussion in Section 5.3 and Appendix B) that by varying $\hat{n} \in \mathbb{N}$ and $f \in \mathcal{L}(D^{\hat{n}})$, we can let $p \coloneqq \mathrm{Mn}^{\hat{n}}_D(f)$ range over all polynomials on Σ_D, and that by varying $\check{n} \in \mathbb{N}$ and $\check{\boldsymbol{x}} \in A^{\check{n}}$, we can let $\check{\boldsymbol{m}} \coloneqq \boldsymbol{T}(\check{\boldsymbol{x}})$ range over all count vectors in $\mathcal{N}_A$. ☐

Proof of Theorem 8. Let, for ease of notation $\Phi \coloneqq \inf_{i \in I} \Phi_i$, then Φ is coherent using Equation (25). Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then, using the representation insensitivity of the coherent Φ_i and Theorem 7:

$$
\begin{aligned}
(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) &\Leftrightarrow (\forall i \in I)(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi_i(A) \\
&\Leftrightarrow (\forall i \in I) p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \Phi_i(D) \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \Phi(D),
\end{aligned}
$$

and this concludes the proof. ☐

Proposition 40. *For all* $\boldsymbol{x} \in A^n$, $\boldsymbol{T}(\boldsymbol{x}{\downarrow_B}) = r_B(\boldsymbol{T}(\boldsymbol{x}))$.

Proof of Proposition 40. Immediate, since $\boldsymbol{x}{\downarrow_B}$ is a sample whose components all belong to B, and for each category in B, the number of times it occurs in $\boldsymbol{x}{\downarrow_B}$ is exactly the same as the number of times it occurs in $\boldsymbol{x}$. $\square$

Proof of Proposition 9. Consider any $\boldsymbol{\vartheta} \in \Sigma_B$, and let, for simplicity of notation $\boldsymbol{\theta} = i_A(\boldsymbol{\vartheta})$. Then since for any $\boldsymbol{n} \in \mathcal{N}_B^n$, with $n := \deg(p) + r$

$$\mathrm{B}_{A,i_A(\boldsymbol{n})}(\boldsymbol{\theta}) = \binom{n}{i_A(\boldsymbol{n})} \prod_{x\in A} \theta_x^{i_A(\boldsymbol{n})_x} = \binom{n}{\boldsymbol{n}} \prod_{x\in B} \vartheta_x^{n_x} = \mathrm{B}_{B,\boldsymbol{n}}(\boldsymbol{\vartheta}),$$

we see that indeed:

$$\mathrm{I}_{B,A}^r(p|\boldsymbol{\theta}) = \sum_{\boldsymbol{n}\in\mathcal{N}_B^n} b_p^n(\boldsymbol{n})\mathrm{B}_{A,i_A(\boldsymbol{n})}(\boldsymbol{\theta}) = \sum_{\boldsymbol{n}\in\mathcal{N}_B^n} b_p^n(\boldsymbol{n})\mathrm{B}_{B,\boldsymbol{n}}(\boldsymbol{\vartheta}) = p(\boldsymbol{\vartheta}). \qquad \square$$

Proof of Proposition 10. When $\deg(p) + r = 0$, then $r = 0$ and $p = c \in \mathbb{R}$, and trivially $\mathrm{I}_{B,A}^r(p|\boldsymbol{\theta}) = \mathrm{I}_{B,A}^0(c)(\boldsymbol{\theta}) = c$. So let us assume that $\deg(p) + r > 0$. First of all, observe that for all $\boldsymbol{n} \in \mathcal{N}_B^{\deg(p)+r}$ and all $\boldsymbol{\theta} \in \Sigma_A$:

$$\begin{aligned}\mathrm{B}_{A,i_A(\boldsymbol{n})}(\boldsymbol{\theta}) &= \binom{\deg(p)+r}{i_A(\boldsymbol{n})} \prod_{x\in A} \theta_x^{i_A(\boldsymbol{n})_x} = \binom{\deg(p)+r}{\boldsymbol{n}} \prod_{x\in B} \theta_x^{n_x} \\ &= \begin{cases} \theta_B^{\deg(p)+r}\mathrm{B}_{B,\boldsymbol{n}}(\boldsymbol{\theta}|_B^+) & \text{if } \theta_B > 0 \\ 0 & \text{otherwise.}\end{cases} \end{aligned} \tag{88}$$

It therefore already follows from Condition (29) that $\mathrm{I}_{B,A}^r(p|\boldsymbol{\theta}) = 0$ if $\theta_B = 0$. Let us therefore assume that $\theta_B > 0$. Then Condition (29) and Equation (88) tell us that:

$$\begin{aligned}\mathrm{I}_{B,A}^r(p|\boldsymbol{\theta}) &= \sum_{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}} b_p^{\deg(p)+r}(\boldsymbol{n})\mathrm{B}_{A,i_A(\boldsymbol{n})}(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}} b_p^{\deg(p)+r}(\boldsymbol{n})\theta_B^{\deg(p)+r}\mathrm{B}_{B,\boldsymbol{n}}(\boldsymbol{\theta}|_B^+) \\ &= \theta_B^{\deg(p)+r} \sum_{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}} b_p^{\deg(p)+r}(\boldsymbol{n})\mathrm{B}_{B,\boldsymbol{n}}(\boldsymbol{\theta}|_B^+) = \theta_B^{\deg(p)+r} p(\boldsymbol{\theta}|_B^+),\end{aligned}$$

which concludes the proof. $\square$

Proposition 41. *For all* $n \in \mathbb{N}$ *and all gambles* f *on* B^n*:*

$$\mathrm{Mn}_A^n(f\mathbb{I}_{B^n}) = \mathrm{I}_{B,A}^r(\mathrm{Mn}_B^n(f)), \text{ where } r := n - \deg(\mathrm{Mn}_B^n(f)).$$

Proof of Proposition 41. First of all, we have for any count vector $\boldsymbol{m}$ in $\mathcal{N}_A^n$ that—with some slight abuse of notation:

$$\mathrm{Hy}_A^n(f\mathbb{I}_{B^n}|\boldsymbol{m}) = \frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x}\in[\boldsymbol{m}]} (f\mathbb{I}_{B^n})(\boldsymbol{x}) = \frac{1}{\nu(\boldsymbol{m})} \sum_{\boldsymbol{x}\in[\boldsymbol{m}]\cap B^n} f(\boldsymbol{x})$$

is zero unless $\boldsymbol{m} = i_A(\boldsymbol{n})$ for some $\boldsymbol{n} \in \mathcal{N}_B^n$. In that case, since then obviously $\nu(\boldsymbol{m}) = \nu(\boldsymbol{n})$, and $\boldsymbol{x} \in [i_A(\boldsymbol{n})] \cap B^n \Leftrightarrow \boldsymbol{x} \in [\boldsymbol{n}]$—again with some slight abuse of notation:

$$\mathrm{Hy}_A^n(f\mathbb{I}_{B^n}|i_A(\boldsymbol{n})) = \frac{1}{\nu(\boldsymbol{n})} \sum_{\boldsymbol{x} \in [\boldsymbol{n}]} f(\boldsymbol{x}) = \mathrm{Hy}_B^n(f|\boldsymbol{n}).$$

Therefore, if we recall Condition (29):

$$\begin{aligned}\mathrm{Mn}_A^n(f\mathbb{I}_{B^n}) &= \sum_{\boldsymbol{n} \in \mathcal{N}_B^n} \mathrm{Hy}_A^n(f\mathbb{I}_{B^n}|i_A(\boldsymbol{n}))\mathrm{B}_{A,i_A(\boldsymbol{n})} = \sum_{\boldsymbol{n} \in \mathcal{N}_B^n} \mathrm{Hy}_B^n(f|\boldsymbol{n})\mathrm{B}_{A,i_A(\boldsymbol{n})} \\ &= \mathrm{I}_{B,A}^r(\mathrm{Mn}_B^n(f)),\end{aligned}$$

where $r := n - \deg(\mathrm{Mn}_B^n(f))$. ∎

Proof of Theorem 11. Fix any category sets A and B such that $B \subseteq A$, any $\check{n}, \hat{n} \in \mathbb{N}$, any $\check{\boldsymbol{x}} \in A^{\check{n}}$ and any gamble f on $B^{\hat{n}}$. We use the notation $\mathcal{H}_A := \Phi(A)$ and $\mathcal{H}_B := \Phi(B)$, and we transform Condition (SP2) using the equivalence in Condition (18). On the one hand, letting $\check{\boldsymbol{m}} := \boldsymbol{T}(\check{\boldsymbol{x}})$ and $r := \hat{n} - \deg(\mathrm{Mn}_B^{\hat{n}}(f))$:

$$\begin{aligned}f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}} &\Leftrightarrow \mathrm{Mn}_A^{\hat{n}}(f\mathbb{I}_{B^{\hat{n}}}) \in \mathcal{H}_A \Leftrightarrow \mathrm{I}_{B,A}^r(\mathrm{Mn}_B^{\hat{n}}(f)) \in \mathcal{H}_A \\ f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} &\Leftrightarrow \mathrm{B}_{A,\check{\boldsymbol{m}}}\,\mathrm{Mn}_A^{\hat{n}}(f\mathbb{I}_{B^{\hat{n}}}) \in \mathcal{H}_A \Leftrightarrow \mathrm{B}_{A,\check{\boldsymbol{m}}}\mathrm{I}_{B,A}^r(\mathrm{Mn}_B^{\hat{n}}(f)) \in \mathcal{H}_A,\end{aligned}$$

where the second equivalences follow from Proposition 41. On the other hand, recalling that $\boldsymbol{T}(\check{\boldsymbol{x}}\downarrow_B) = r_B(\boldsymbol{T}(\check{\boldsymbol{x}})) = r_B(\check{\boldsymbol{m}})$ by Proposition 40:

$$\begin{aligned}f \in \mathcal{D}_B^{\hat{n}} &\Leftrightarrow \mathrm{Mn}_B^{\hat{n}}(f) \in \mathcal{H}_B \\ f \in \mathcal{D}_B^{\hat{n}}\rfloor\check{\boldsymbol{x}}\downarrow_B &\Leftrightarrow \mathrm{B}_{B,r_B(\check{\boldsymbol{m}})}\,\mathrm{Mn}_B^{\hat{n}}(f) \in \mathcal{H}_B.\end{aligned}$$

This tells us that the equivalences in Condition (SP2) can be rewritten as:

$$\begin{aligned}\mathrm{I}_{B,A}^r(\mathrm{Mn}_B^{\hat{n}}(f)) \in \mathcal{H}_A &\Leftrightarrow \mathrm{Mn}_B^{\hat{n}}(f) \in \mathcal{H}_B \\ \mathrm{B}_{A,\check{\boldsymbol{m}}}\mathrm{I}_{B,A}^r(\mathrm{Mn}_B^{\hat{n}}(f)) \in \mathcal{H}_A &\Leftrightarrow \mathrm{B}_{B,r_B(\check{\boldsymbol{m}})}\,\mathrm{Mn}_B^{\hat{n}}(f) \in \mathcal{H}_B.\end{aligned}$$

The proof is complete if we recall from the discussion in Section 5.3 and Appendix B that by varying $\hat{n} \in \mathbb{N}$ and $f \in \mathcal{L}(B^{\hat{n}})$, we can let $p := \mathrm{Mn}_B^{\hat{n}}(f) = \mathrm{CoMn}_B^{\hat{n}}(\mathrm{Hy}_B^{\hat{n}}(f))$ range over all polynomials on Σ_B and $r = \hat{n} - \deg(\mathrm{Mn}_B^{\hat{n}}(f))$ range over all elements of $\mathbb{N}_0$, and that by varying $\check{n} \in \mathbb{N}$ and $\check{\boldsymbol{x}} \in A^{\check{n}}$, we can let $\check{\boldsymbol{m}} := \boldsymbol{T}(\check{\boldsymbol{x}})$ range over all count vectors in $\mathcal{N}_A$. ∎

Proof of Theorem 12. Let, for ease of notation $\Phi := \inf_{i \in I} \Phi_i$, then Φ is coherent using Equation (25). Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $r \in \mathbb{N}_0$. Then, using the specificity of the Φ_i:

$$\begin{aligned}\mathrm{I}_{B,A}^r(p)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) &\Leftrightarrow (\forall i \in I)\mathrm{I}_{B,A}^r(p)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi_i(A) \\ &\Leftrightarrow (\forall i \in I)p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \Phi_i(B) \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \Phi(B),\end{aligned}$$

which concludes the proof. ∎

E.3 Proofs of Results in Section 8

Proof of Proposition 13. For sufficiency, fix a category set A, a gamble g on A, and a count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$. Condition (RI4) with $D := g(A)$, $\rho := g$, $f := \mathrm{id}_D$ yields Condition (RI5).

For necessity, fix category sets A and D such that there is an onto map $\rho\colon A \to D$, a gamble f on D, and a count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$. Observe that $(f \circ \rho)(A) = f(D)$ and that for all $r \in f(D)$

$$\begin{aligned} R_{f\circ\rho}(\boldsymbol{m})_r = \sum_{x\in A\colon (f\circ\rho)(x)=r} m_x &= \sum_{z\in D\colon f(z)=r}\ \sum_{x\in A\colon \rho(x)=z} m_x \\ &= \sum_{z\in D\colon f(z)=r} R_\rho(\boldsymbol{m})_z = R_f(R_\rho(\boldsymbol{m}))_r, \end{aligned}$$

so $R_{f\circ\rho} = R_f \circ R_\rho$. We now infer by invoking Condition (RI5) twice that:

$$\begin{aligned} f\circ\rho \in \mathcal{D}^1_A\rfloor\boldsymbol{m} &\Leftrightarrow \mathrm{id}_{(f\circ\rho)(A)} \in \mathcal{D}^1_{(f\circ\rho)(A)}\rfloor R_{f\circ\rho}(\boldsymbol{m}) \\ &\Leftrightarrow \mathrm{id}_{f(D)} \in \mathcal{D}^1_{f(D)}\rfloor R_f(R_\rho(\boldsymbol{m})) \Leftrightarrow f \in \mathcal{D}^1_D\rfloor R_\rho(\boldsymbol{m}), \end{aligned}$$

concluding the proof. □

Proof of Theorem 14. The arguments in this proof rely heavily on the following expression for the lower probability function:

$$\begin{aligned} \varphi(n,k) &= \sup\left\{\alpha\in\mathbb{R}\colon \mathbb{I}_{\{a\}} - \alpha \in \mathcal{D}^1_{\{a,b\}}\rfloor(k,n-k)\right\} \\ &= \sup\left\{\alpha\in\mathbb{R}\colon \theta_a^k\theta_b^{n-k}[\theta_a-\alpha] \in \Phi(\{a,b\})\right\} \end{aligned} \tag{89}$$

and other related expressions that are equivalent to it by representation insensitivity and Bernstein coherence [B3]. Both expressions follow from Equations (32) and (33), Bernstein coherence [B3] and representation insensitivity in its form (RI4).

L1. Immediate from Bernstein coherence and the fact that $\varphi(n,k)$ is a lower probability: use Equation (89), B2 and B4.

L2. Fix any non-negative integers n, k and ℓ such that $k+\ell \leq n$. Consider any real $\alpha < \varphi(n,k)$ and $\beta < \varphi(n,\ell)$, then it follows from applying Equation (89) and Condition (RI4) that both $\theta_x^k\theta_y^\ell\theta_z^{n-k-\ell}[\theta_x-\alpha] \in \Phi(\{x,y,z\})$ and $\theta_x^k\theta_y^\ell\theta_z^{n-k-\ell}[\theta_y-\beta] \in \Phi(\{x,y,z\})$, whence, by Bernstein coherence [B3], $\theta_x^k\theta_y^\ell\theta_z^{n-k-\ell}[(\theta_x+\theta_y)-(\alpha+\beta)] \in \Phi(\{x,y,z\})$. Applying Equation (89) and Condition (RI4) again tells us that $\theta_u^{k+\ell}\theta_z^{n-k-\ell}[\theta_u-(\alpha+\beta)] \in \Phi(\{u,z\})$, whence $\alpha+\beta \leq \varphi(n,k+\ell)$.

L3, L4 and L5 are immediate consequences of L1 and L2.

L6. Consider the category set $A := \{a,b\}$ and the count vector $\boldsymbol{m}$ with $m_a := k$ and $m_b := n-k$. Define the gamble g on A by $g(a) := \varphi(n+1,k+1)$ and $g(b) := \varphi(n+1,k)$. Then $g(a) \geq g(b)$ by L5, and therefore the coherence [P5 and P3] of the predictive lower prevision $\underline{P}^1_A(\cdot|\boldsymbol{m})$ tells us that $\underline{P}^1_A(g|\boldsymbol{m}) = g(b) + [g(a)-g(b)]\underline{P}^1_A(\{a\}|\boldsymbol{m}) = \varphi(n+1,k) + \varphi(n,k)[\varphi(n+1,k+1)-\varphi(n+1,k)]$ [see also Equation (33)]. So it clearly suffices to prove that $\underline{P}^1_A(g|\boldsymbol{m}) \leq \varphi(n,k) = \underline{P}^1_A(\{a\}|\boldsymbol{m})$. Consider any $\alpha < \underline{P}^1_A(g|\boldsymbol{m})$, then it follows using Equation (31) that:

$$\theta_a^k\theta_b^{n-k}[g(a)\theta_a + g(b)\theta_b - \alpha] \in \Phi(A). \tag{90}$$

Also, for any $\epsilon > 0$, both $\theta_a^{k+1}\theta_b^{n-k}[\theta_a - g(a) + \epsilon] \in \Phi(A)$ and $\theta_a^k\theta_b^{n+1-k}[\theta_a - g(b) + \epsilon] \in \Phi(A)$, and therefore, by coherence [B3], and recalling that $\theta_a + \theta_b = 1$,

$$\begin{aligned}\Phi(A) \ni \theta_a^{k+1}\theta_b^{n-k}[\theta_a - g(a) + \epsilon] + \theta_a^k\theta_b^{n+1-k}[\theta_a - g(b) + \epsilon]&\\ = \theta_a^k\theta_b^{n-k}[\theta_a - g(a)\theta_a - g(b)\theta_b + \epsilon].&\end{aligned} \tag{91}$$

Combining Statements (90) and (91) using coherence [B3], this leads to $\theta_a^k\theta_b^{n-k}[\theta_a - \alpha + \epsilon] \in \Phi(A)$, whence $\varphi(n,k) \geq \alpha - \epsilon$, and this completes the proof.

L7. Use L1 and L5 to find that $\varphi(n,k)[\varphi(n+1,k+1) - \varphi(n+1,k)] \geq 0$, and then use L6.

L8. That $s_n \geq 0$ follows from L4, so we only need to prove that $s_{n+1} \geq s_n$, or equivalently, that $\varphi(n,1) \geq \varphi(n+1,1)[1 + \varphi(n,1)]$. Indeed:

$$\begin{aligned}\varphi(n,1) &\geq \varphi(n+1,1) + \varphi(n,1)[\varphi(n+1,2) - \varphi(n+1,1)]\\ &\geq \varphi(n+1,1) + \varphi(n,1)[2\varphi(n+1,1) - \varphi(n+1,1)]\\ &= \varphi(n+1,1) + \varphi(n,1)\varphi(n+1,1),\end{aligned}$$

where the first inequality follows from L6 with $k = 1$, and the second from L4 and L1.

L9. The inequalities hold trivially for $n = 0$, due to L1. So consider any $n \in \mathbb{N}$, and category sets $A \coloneqq \{x, y\}$ and $B \coloneqq \{x_1, x_2, \dots, x_n, y\}$. Let $0 < \epsilon < \alpha$ and $\beta \coloneqq \alpha - \epsilon > 0$. Since $\varphi(1,1) > \beta$, we see that $\theta_x[\theta_x - \beta] \in \Phi(A)$, or equivalently, $\theta_x[\theta_x(1-\beta) - \beta\theta_y] \in \Phi(A)$, since $\theta_x + \theta_y = 1$. Representation insensitivity [use Equation (89) and Condition (RI4)] then tells us that $\vartheta_{x_k}[\vartheta_{x_k}(1-\beta) - \beta\vartheta_y] \in \Phi(\{x_k, y\})$, and specificity [use Theorem 11] allows us to infer from this that $(\prod_{k=1}^n \vartheta_{x_k})[\vartheta_{x_k}(1-\beta) - \beta\vartheta_y] \in \Phi(B)$, for all $k \in \{1, \dots, n\}$. Now infer from coherence [B3] that $(\prod_{k=1}^n \vartheta_{x_k})[\sum_{k=1}^n \vartheta_{x_k}(1-\beta) - n\beta\vartheta_y] \in \Phi(B)$, and apply representation insensitivity to get to $\theta_x^n[\theta_x(1-\beta) - n\beta\theta_y] \in \Phi(A)$. Since $\theta_y = 1 - \theta_x$, this is equivalent to $\theta_x^n[\theta_x(1-\beta+n\beta) - n\beta] \in \Phi(A)$. This shows that $\phi(n,n) \geq \frac{n\beta}{1-\beta+n\beta}$, using Equation (89). The rest of the proof is now immediate. ☐

E.4 Proofs of Results in Section 9

Proof of Theorem 15. That Φ_{V} is coherent is obvious, because for each category set $A \in \mathbb{F}$, $\Phi_{\mathrm{V}}(A) = \mathcal{V}^+(A)$ is a Bernstein coherent set of polynomials on Σ_A.

To prove representation insensitivity, we use Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then indeed

$$(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^+(A) \Leftrightarrow p \circ R_\rho \in \mathcal{V}^+(A) \Leftrightarrow p \in \mathcal{V}^+(D) \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \mathcal{V}^+(D),$$

where the first and last equivalences follow from Proposition 30, and the second one from Lemma 48 with $K = A$. ☐

E.5 Proofs of Results in Section 10

Proof of Theorem 16. That Φ_{V} is coherent is obvious, because for each category set $A \in \mathbb{F}$, $\Phi_{\mathrm{V}}(A) = \mathcal{V}^{++}(A)$ is obviously a convex cone that includes $\mathcal{V}^{+}(A)$ [Proposition 28] and does not contain the zero polynomial: $\mathcal{V}^{++}(A)$ is therefore a Bernstein coherent set of polynomials on Σ_A.

To prove representation insensitivity, we use Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then indeed

$$\begin{aligned}(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^{++}(A) &\Leftrightarrow (\forall \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A))p(R_\rho(\boldsymbol{\theta}))\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A))p(R_\rho(\boldsymbol{\theta})) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\vartheta} \in \mathrm{int}(\Sigma_D))p(\boldsymbol{\vartheta}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\vartheta} \in \mathrm{int}(\Sigma_D))p(\boldsymbol{\vartheta})\mathrm{B}_{D,R_\rho(\boldsymbol{m})}(\boldsymbol{\vartheta}) > 0 \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \mathcal{V}^{++}(D),\end{aligned}$$

where the second and fourth equivalences follow from the Bernstein positivity of the Bernstein basis polynomials and Proposition 28, and the third one from Lemma 48 with $K = A$.

To prove specificity, we use Theorem 11. Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $r \in \mathbb{N}_0$. Then indeed:

$$\begin{aligned}\mathrm{I}^r_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{V}^{++}(A) &\Leftrightarrow (\forall \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A))\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta})\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\theta} \in \mathrm{int}(\Sigma_A))\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\vartheta} \in \mathrm{int}(\Sigma_B))p(\boldsymbol{\vartheta}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\vartheta} \in \mathrm{int}(\Sigma_B))p(\boldsymbol{\vartheta})\mathrm{B}_{B,r_B(\boldsymbol{m})}(\boldsymbol{\vartheta}) > 0 \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \mathcal{V}^{++}(B),\end{aligned}$$

where the second and fourth equivalences follow from the Bernstein positivity of the Bernstein basis polynomials and Proposition 28, and the third one from Lemma 52 with $K = A$. □

E.6 Proofs of Results in Section 11

Below, we use the convenient device of identifying, for any proper subset B of A, an element $\boldsymbol{\vartheta}$ of Σ_B with the unique corresponding element $\boldsymbol{\theta} = i_A(\boldsymbol{\vartheta})$ of Σ_A whose components outside B are zero:

$$(\forall x \in B)\theta_x = \vartheta_x \text{ and } (\forall x \in A \setminus B)\theta_x = 0.$$

Also observe that, using this convention, we can identify $\mathrm{int}(\Sigma_{A[\boldsymbol{m}]})$ with a subset of Σ_A, and then characterise it as follows:

$$\text{for any } \boldsymbol{\theta} \in \Sigma_A\colon \boldsymbol{\theta} \in \mathrm{int}(\Sigma_{A[\boldsymbol{m}]}) \Leftrightarrow (\forall x \in A)(\theta_x > 0 \Leftrightarrow m_x > 0).$$

Proof of Proposition 17. It clearly suffices to prove that $\mathcal{V}^{+}(A) \subseteq \mathcal{H}_{\mathrm{SC},A}$ and $0 \notin \mathcal{H}_{\mathrm{SC},A}$.

The first statement is easy to prove because $\mathcal{A}_{\mathrm{SC},A}$ trivially includes all non-constant Bernstein basis polynomials, by Proposition 28. Since $\mathcal{V}^{+}(A)$ consists of finite, strictly positive linear combinations of these non-constant Bernstein basis polynomials, we immediately have that $\mathcal{V}^{+}(A) \subseteq \mathcal{H}_{\mathrm{SC},A}$.

To prove the second statement, suppose *ex absurdo* that $0 \in \mathcal{H}_{\mathrm{SC},A}$. This implies that there are finitely many $n_k > 0$, count vectors $\boldsymbol{m}_k$ in $\mathcal{N}_A^{n_k}$ and $p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}_k]}(A)$ such that $0 = \sum_k p_k \mathrm{B}_{A,\boldsymbol{m}_k}$. It is always possible to find (at least) one such count vector, $\boldsymbol{m}_1$ say, for which $A[\boldsymbol{m}_k] \not\subset A[\boldsymbol{m}_1]$ for all k. In other words, we have either $A[\boldsymbol{m}_k] = A[\boldsymbol{m}_1]$ or $A[\boldsymbol{m}_k] \setminus A[\boldsymbol{m}_1] \neq \emptyset$. Now consider any $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_{A[\boldsymbol{m}_1]})$. If $A[\boldsymbol{m}_k] \setminus A[\boldsymbol{m}_1] \neq \emptyset$, then $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) = 0$. If $A[\boldsymbol{m}_k] = A[\boldsymbol{m}_1]$, then $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) > 0$, and moreover, since $p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}_k]}(A)$, $p_k(\boldsymbol{\theta}) > 0$. Hence $0 = \sum_k p_k(\boldsymbol{\theta}) \mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) > 0$, a contradiction. □

Lemma 42. *Consider any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and $p \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$, so there are $\ell \in \mathbb{N}$, $n_k \in \mathbb{N}_0$ such that $m_A + n_k > 0$, $\boldsymbol{m}_k \in \mathcal{N}_A^{n_k}$ and $p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}+\boldsymbol{m}_k]}(A)$ such that $p = \sum_{k=1}^{\ell} p_k \mathrm{B}_{A,\boldsymbol{m}_k}$. Then*

$$S_{A,\boldsymbol{m}}(p) = \{K \subseteq A \colon A[\boldsymbol{m}+\boldsymbol{m}_k] \subseteq K \text{ for some } k \in \{1,\dots,\ell\}\}$$

and therefore

$$\min S_{A,\boldsymbol{m}}(p) = \min\{A[\boldsymbol{m}+\boldsymbol{m}_k] \colon k \in \{1,\dots,\ell\}\}.$$

Proof of Lemma 42. The second statement is trivial, given the first. So we restrict our attention to proving the first statement.

Assume first that $A[\boldsymbol{m}+\boldsymbol{m}_r] \subseteq K \subseteq A$ for some $r \in \{1,\dots,\ell\}$. Then clearly $K \neq \emptyset$, since $m_A + n_r > 0$. We may assume without loss of generality that $A[\boldsymbol{m}+\boldsymbol{m}_r]$ is a minimal element of the set $\{A[\boldsymbol{m}+\boldsymbol{m}_k] \colon k \in \{1,\dots,\ell\}\}$. Consider any $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_{A[\boldsymbol{m}+\boldsymbol{m}_r]})$, whence also $\boldsymbol{\theta} \in \Sigma_K$. Now for all $k \in \{1,\dots,\ell\}$ such that $A[\boldsymbol{m}+\boldsymbol{m}_k] = A[\boldsymbol{m}+\boldsymbol{m}_r]$—and there clearly is at least one such k—we see that both $p_k(\boldsymbol{\theta}) > 0$ since $p_k \in \mathcal{V}^{++}_{[\boldsymbol{m}+\boldsymbol{m}_k]}(A)$, and $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) > 0$, whence $(p_k \mathrm{B}_{A,\boldsymbol{m}_k})(\boldsymbol{\theta}) > 0$. For all other k we must have that $A[\boldsymbol{m}+\boldsymbol{m}_k] \setminus A[\boldsymbol{m}+\boldsymbol{m}_r] \neq \emptyset$, and therefore $(p_k \mathrm{B}_{A,\boldsymbol{m}_k})(\boldsymbol{\theta}) = 0$ since $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) = 0$. This guarantees that $p(\boldsymbol{\theta}) = \sum_{k=1}^{\ell} (p_k \mathrm{B}_{A,\boldsymbol{m}_k})(\boldsymbol{\theta}) > 0$, whence indeed $K \in S_{A,\boldsymbol{m}}(p)$, since we already know that $\boldsymbol{\theta} \in \Sigma_K$, $A[\boldsymbol{m}] \subseteq A[\boldsymbol{m}+\boldsymbol{m}_r] \subseteq K$ and $K \neq \emptyset$.

Assume, conversely, that $K \in S_{A,\boldsymbol{m}}(p)$, which implies that $\emptyset \neq K \subseteq A$ and $A[\boldsymbol{m}] \subseteq K$, and that there is some $\boldsymbol{\vartheta} \in \Sigma_K$ such that $p(\boldsymbol{\vartheta}) \neq 0$. Observe that $A[\boldsymbol{m}+\boldsymbol{m}_k] = A[\boldsymbol{m}] \cup A[\boldsymbol{m}_k]$, and assume *ex absurdo* that $A[\boldsymbol{m}+\boldsymbol{m}_k] \nsubseteq K$ and therefore $A[\boldsymbol{m}_k] \nsubseteq K$ for all $k \in \{1,\dots,\ell\}$. Fix any $k \in \{1,\dots,\ell\}$, then there is some $x \in A[\boldsymbol{m}_k]$ such that $x \notin K$, and therefore $\vartheta_x = 0$, whence $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\vartheta}) = 0$. This shows that $p(\boldsymbol{\vartheta}) = \sum_{k=1}^{\ell} (p_k \mathrm{B}_{A,\boldsymbol{m}_k})(\boldsymbol{\vartheta}) = 0$, a contradiction. □

Lemma 43. *Consider any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, any $p \in \mathcal{V}(A)$ and any $n \in \mathbb{N}$ such that $n \geq \deg(p)$. Then for all $\boldsymbol{\mu} \in \mathcal{N}_A^n$:*

$$b_p^n(\boldsymbol{\mu}) \neq 0 \Rightarrow (\exists K \in \min S_{A,\boldsymbol{m}}(p)) K \setminus A[\boldsymbol{m}] \subseteq A[\boldsymbol{\mu}].$$

Proof of Lemma 43. Fix any $\boldsymbol{\mu}$ in $\mathcal{N}_A^n$. We prove the contraposition, so suppose that for all K in $\min S_{A,\boldsymbol{m}}(p)$, we have that $K \setminus A[\boldsymbol{m}] \nsubseteq A[\boldsymbol{\mu}]$ and therefore $K \nsubseteq A[\boldsymbol{m}+\boldsymbol{\mu}]$, since $A[\boldsymbol{m}+\boldsymbol{\mu}] = A[\boldsymbol{m}] \cup A[\boldsymbol{\mu}]$. Hence, $A[\boldsymbol{m}+\boldsymbol{\mu}] \notin S_{A,\boldsymbol{m}}(p)$. Since moreover $\emptyset \neq A[\boldsymbol{m}+\boldsymbol{\mu}]$ and $A[\boldsymbol{m}] \subseteq A[\boldsymbol{m}+\boldsymbol{\mu}]$, we infer from Equation (46) that $p_{|\Sigma_B} = 0$, where we let, for ease of notation, $B \coloneqq A[\boldsymbol{m}+\boldsymbol{\mu}]$. We can rewrite this as [see also Lemma 47]:

$$0 = p_{|\Sigma_B} = \sum_{\boldsymbol{\eta} \in \mathcal{N}_A^n} b_p^n(\boldsymbol{\eta}) \mathrm{B}_{A,\boldsymbol{\eta}|\Sigma_B} = \sum_{\boldsymbol{\eta} \in \mathcal{N}_B^n} b_p^n(\boldsymbol{\eta}) \mathrm{B}_{A,\boldsymbol{\eta}|\Sigma_B} = \sum_{\boldsymbol{\eta} \in \mathcal{N}_B^n} b_p^n(\boldsymbol{\eta}) \mathrm{B}_{B,\boldsymbol{\eta}}.$$

Due to the uniqueness of the Bernstein expansion, this is only possible if $b_p^n(\boldsymbol{\eta}) = 0$ for all $\boldsymbol{\eta} \in \mathcal{N}_{A[\boldsymbol{m}+\boldsymbol{\mu}]}^n$. This concludes the proof since, clearly, $\boldsymbol{\mu} \in \mathcal{N}_{A[\boldsymbol{m}+\boldsymbol{\mu}]}^n$. $\square$

Proof of Proposition 18. First, assume that $p \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$, implying that $p = \sum_{k=1}^{\ell} p_k \mathrm{B}_{A,\boldsymbol{m}_k}$ for some $\ell \in \mathbb{N}$, $n_k \in \mathbb{N}_0$ such that $m_A + n_k > 0$, $\boldsymbol{m}_k \in \mathcal{N}_A^{n_k}$ and $p_k \in \mathcal{V}_{[\boldsymbol{m}+\boldsymbol{m}_k]}^{++}(A)$. It already follows from Lemma 42 that $p \neq 0$ and that $\min S_{A,\boldsymbol{m}}(p) = \min \{A[\boldsymbol{m}+\boldsymbol{m}_k] \colon k \in \{1, \dots, \ell\}\}$. Consider now any $K \in \min \{A[\boldsymbol{m}+\boldsymbol{m}_k] \colon k \in \{1, \dots, \ell\}\}$ and any $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_K)$. Then for all k, we have that either $A[\boldsymbol{m}+\boldsymbol{m}_k] = K$ or $A[\boldsymbol{m}+\boldsymbol{m}_k] \setminus K \neq \emptyset$. If $A[\boldsymbol{m}+\boldsymbol{m}_k] = K$—which happens for at least one k, due to our choice of K—then $p_k(\boldsymbol{\theta}) > 0$ and $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) > 0$. If $A[\boldsymbol{m}+\boldsymbol{m}_k] \setminus K \neq \emptyset$, then since $A[\boldsymbol{m}] \subseteq K$, $A[\boldsymbol{m}_k] \setminus K \neq \emptyset$, implying that $\mathrm{B}_{A,\boldsymbol{m}_k}(\boldsymbol{\theta}) = 0$. Hence, $p(\boldsymbol{\theta}) > 0$. Since this holds for all $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_K)$, we find that $p_{|\Sigma_K} \in \mathcal{V}^{++}(K)$.

Assume, conversely, that $p \in \mathcal{V}(A) \setminus \{0\}$ and that $p_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ for all $K \in \min S_{A,\boldsymbol{m}}(p)$. Fix any $n \in \mathbb{N}$ such that $n \geq \deg(p)$, then $p = \sum_{\boldsymbol{\mu} \in \mathcal{N}_A^n} b_p^n(\boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}} = \sum_{\boldsymbol{\mu} \in M} b_p^n(\boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}}$, with $M \coloneqq \left\{\boldsymbol{\mu} \in \mathcal{N}_A^n \colon b_p^n(\boldsymbol{\mu}) \neq 0\right\}$. Since $p \neq 0$, we infer from Equation (46) that $\min S_{A,\boldsymbol{m}}(p) \neq \emptyset$ [observe that $A \in S_{A,\boldsymbol{m}}(p)$]. We know from Lemma 43 that for any $\boldsymbol{\mu} \in M$, there is at least one $K \in \min S_{A,\boldsymbol{m}}(p)$ such that $K \setminus A[\boldsymbol{m}] \subseteq A[\boldsymbol{\mu}]$. Let us just pick any of these K, and call it $K_{\boldsymbol{\mu}}$. Now let, for any $K \in \min S_{A,\boldsymbol{m}}(p)$, $M_K \coloneqq \{\boldsymbol{\mu} \in M \colon K_{\boldsymbol{\mu}} = K\}$, then we have found a way to divide M into disjoint subsets M_K, one for every $K \in \min S_{A,\boldsymbol{m}}(p)$ and some of which may be empty, such that $K \setminus A[\boldsymbol{m}] \subseteq A[\boldsymbol{\mu}]$ for all $\boldsymbol{\mu} \in M_K$, $M = \bigcup_{K \in \min S_{A,\boldsymbol{m}}(p)} M_K$ and therefore $p = \sum_{K \in \min S_{A,\boldsymbol{m}}(p)} \sum_{\boldsymbol{\mu} \in M_K} b_p^n(\boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}}$.

Now fix any $K \in \min S_{A,\boldsymbol{m}}(p)$, then we construct a count vector $\boldsymbol{m}_K$ by letting $(m_K)_x \coloneqq 1$ if $x \in K \setminus A[\boldsymbol{m}]$ and $(m_K)_x \coloneqq 0$ otherwise. Notice that $\boldsymbol{m}_K \in \mathcal{N}_A^{n_K}$, with n_K the number of elements $|K \setminus A[\boldsymbol{m}]|$ in the set $K \setminus A[\boldsymbol{m}]$, and therefore $n_K \leq n$. Consider any $\boldsymbol{\mu} \in M_K$, then since $(m_K)_x = 1$ implies that $x \in A[\boldsymbol{\mu}]$ and therefore $\mu_x \geq 1$, we see that for all $\boldsymbol{\theta} \in \Sigma_A$:

$$\begin{aligned}
\mathrm{B}_{A,\boldsymbol{\mu}}(\boldsymbol{\theta}) = \nu(\boldsymbol{\mu}) \prod_{x \in A[\boldsymbol{\mu}]} \theta_x^{\mu_x} &= \nu(\boldsymbol{\mu}) \prod_{x \in A[\boldsymbol{\mu}]} \theta_x^{\mu_x - (m_K)_x} \prod_{x \in A[\boldsymbol{\mu}]} \theta_x^{(m_K)_x} \\
&= \lambda(K, \boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}-\boldsymbol{m}_K}(\boldsymbol{\theta}) \mathrm{B}_{A,\boldsymbol{m}_K}(\boldsymbol{\theta}),
\end{aligned}$$

where $\lambda(K, \boldsymbol{\mu}) \coloneqq \nu(\boldsymbol{\mu}) \nu(\boldsymbol{\mu} - \boldsymbol{m}_K)^{-1} \nu(\boldsymbol{m}_K)^{-1}$. Hence, we can rewrite $\sum_{\boldsymbol{\mu} \in M_K} b_p^n(\boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}}$ as $p_K \mathrm{B}_{A,\boldsymbol{m}_K}$, where $p_K \coloneqq \sum_{\boldsymbol{\mu} \in M_K} \lambda(K, \boldsymbol{\mu}) b_p^n(\boldsymbol{\mu}) \mathrm{B}_{A,\boldsymbol{\mu}-\boldsymbol{m}_K}$. In this way, we find that $p = \sum_{K \in \min S_{A,\boldsymbol{m}}(p)} \mathrm{B}_{A,\boldsymbol{m}_K} p_K$.

Hence, if we fix any $K \in \min S_{A,\boldsymbol{m}}(p) \neq \emptyset$, then we are left to prove that $m_A + n_K > 0$ and $p_K \in \mathcal{V}_{[\boldsymbol{m}+\boldsymbol{m}_K]}^{++}(A)$. Assume first, *ex absurdo*, that $m_A + n_K = 0$. Then in particular $K = \emptyset$, which contradicts $K \in S_{A,\boldsymbol{m}}(p)$. So it remains to prove that $p_K \in \mathcal{V}_{[\boldsymbol{m}+\boldsymbol{m}_K]}^{++}(A)$. Consider any $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_{A[\boldsymbol{m}+\boldsymbol{m}_K]})$. Then we can derive from $K \in \min S_{A,\boldsymbol{m}}(p) \subseteq S_{A,\boldsymbol{m}}(p)$ that $A[\boldsymbol{m}] \subseteq K$. Since $A[\boldsymbol{m}_K] = K \setminus A[\boldsymbol{m}]$, this implies that $A[\boldsymbol{m}+\boldsymbol{m}_K] = A[\boldsymbol{m}] \cup A[\boldsymbol{m}_K] = A[\boldsymbol{m}] \cup (K \setminus A[\boldsymbol{m}]) = K$, and therefore also $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_K)$. For all $K' \in \min S_{A,\boldsymbol{m}}(p) \setminus \{K\}$, $K' \setminus K \neq \emptyset$ and therefore $\mathrm{B}_{A,\boldsymbol{m}_{K'}}(\boldsymbol{\theta}) = 0$. Hence, $p(\boldsymbol{\theta}) = \mathrm{B}_{A,\boldsymbol{m}_K}(\boldsymbol{\theta}) p_K(\boldsymbol{\theta})$. We know that $p(\boldsymbol{\theta}) > 0$ because $p_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ and that $\mathrm{B}_{A,\boldsymbol{m}_K}(\boldsymbol{\theta}) > 0$ because $A[\boldsymbol{m}_K] = K \setminus A[\boldsymbol{m}] \subseteq K$. We conclude that indeed $p_K(\boldsymbol{\theta}) > 0$. $\square$

Lemma 44. *For all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and $p \in \mathcal{V}(A)$:*

$$S_{A,\boldsymbol{m}}(p) = S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}}) \text{ and therefore } \min S_{A,\boldsymbol{m}}(p) = \min S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}}).$$

Proof of Lemma 44. First, assume that $K \in S_{A,\boldsymbol{m}}(p)$. Then $\emptyset \neq K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $p_{|\Sigma_K} \neq 0$. From this last inequality and the continuity of polynomials, we infer that there is some $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$ such that $p(\boldsymbol{\theta}) \neq 0$. Since $A[\boldsymbol{m}] \subseteq K$, we find that $p(\boldsymbol{\theta})\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \neq 0$ and therefore $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \neq 0$.

Assume, conversely, that $K \in S_{A,\mathbf{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$. Then $\emptyset \neq K \subseteq A$ and $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \neq 0$. This last inequality implies that there is some $\boldsymbol{\theta} \in \Sigma_K$ such that $(p\mathrm{B}_{A,\boldsymbol{m}})(\boldsymbol{\theta}) \neq 0$ and therefore both $\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \neq 0$ and $p(\boldsymbol{\theta}) \neq 0$. From $\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) \neq 0$, we derive that $A[\boldsymbol{m}] \subseteq K$ and from $p(\boldsymbol{\theta}) \neq 0$, we derive that $p_{|\Sigma_K} \neq 0$. ☐

Proof of Proposition 19. By the way $\mathcal{H}_{\mathrm{SC},A}$ and $\mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$ are constructed [see the defining expressions (43) and (44)], it clearly suffices to prove that $\mathcal{H}_{\mathrm{SC},A}\rfloor\boldsymbol{m} \subseteq \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$. Consider therefore any $p \in \mathcal{V}(A)$ such that $p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}_{\mathrm{SC},A}$, which by Proposition 18, implies that $p\mathrm{B}_{A,\boldsymbol{m}} \neq 0$ and that $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ for all $K \in \min S_{A,\mathbf{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$. We now set out to prove that $p \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$. Applying Proposition 18 again, and since, clearly, $p \neq 0$, we see that it suffices to show that $p_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ for all $K \in \min S_{A,\boldsymbol{m}}(p)$. So consider any $K \in \min S_{A,\boldsymbol{m}}(p)$. Then, by Lemma 44, $K \in \min S_{A,\mathbf{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$, so we have already argued above that $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{V}^{++}(K)$. Hence indeed also $p_{|\Sigma_K} \in \mathcal{V}^{++}(K)$. ☐

Proof of Equation (48). Combining Equations (47) and (30), we see that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^1_{\mathrm{SC},A}\rfloor\check{\boldsymbol{m}} = \{f \in \mathcal{L}(A) \colon \mathrm{S}_A(f) \in \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}\}\,. \tag{92}$$

Also, for any $f \in \mathcal{L}(A)$ and any $\emptyset \neq K \subseteq A$:

$$\mathrm{S}_A(f) = 0 \Leftrightarrow f = 0, \mathrm{S}_A(f)_{|\Sigma_K} \in \mathcal{V}^{++}(K) \Leftrightarrow f_{|K} > 0 \text{ and } \mathrm{S}_A(f)_{|\Sigma_K} = 0 \Leftrightarrow f_{|K} = 0. \tag{93}$$

We start with the case $\check{\boldsymbol{m}} = \mathbf{0}$. For any $f \in \mathcal{L}(A)$:

$$\min S_{A,\mathbf{0}}(\mathrm{S}_A(f)) = \{\{x\} \colon x \in A \text{ and } f(x) \neq 0\}\,,$$

because of Statement (93). Hence, by Proposition 18 and Equations (92) and (93): $\mathcal{D}^1_{\mathrm{SC},A} = \mathcal{L}_{>0}(A)$.

Next, we consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A$. For all $f \in \mathcal{L}(A)$:

$$\min S_{A,\check{\boldsymbol{m}}}(\mathrm{S}_A(f)) = \begin{cases} \{A[\check{\boldsymbol{m}}]\} & \text{if } f_{|A[\check{\boldsymbol{m}}]} \neq 0 \\ \{A[\check{\boldsymbol{m}}] \cup \{x\} \colon x \in A \setminus A[\check{\boldsymbol{m}}] \text{ and } f(x) \neq 0\} & \text{if } f_{|A[\check{\boldsymbol{m}}]} = 0 \end{cases} \tag{94}$$

because of Equation (93). Now recall Proposition 18 and Equations (92) and (93) and consider two cases: $f_{|A[\check{\boldsymbol{m}}]} \neq 0$ and $f_{|A[\check{\boldsymbol{m}}]} = 0$. If $f_{|A[\check{\boldsymbol{m}}]} \neq 0$, then $f \in \mathcal{D}^1_{\mathrm{SC},A}\rfloor\check{\boldsymbol{m}}$ if and only if $f \neq 0$ [which is redundant] and $f_{|A[\check{\boldsymbol{m}}]} > 0$ or, equivalently [since $f_{|A[\check{\boldsymbol{m}}]} \neq 0$], if $f \in \left\{h \in \mathcal{L}(A) \colon h_{|A[\check{\boldsymbol{m}}]} > 0\right\} \cup \mathcal{L}_{>0}(A)$. If $f_{|A[\check{\boldsymbol{m}}]} = 0$, then $f \in \mathcal{D}^1_{\mathrm{SC},A}\rfloor\check{\boldsymbol{m}}$ if and only if $f \neq 0$ and $f(x) \geq 0$ for all $x \in A \setminus A[\check{\boldsymbol{m}}]$ or, equivalently [since $f_{|A[\check{\boldsymbol{m}}]} = 0$], again if $f \in \left\{h \in \mathcal{L}(A) \colon h_{|A[\check{\boldsymbol{m}}]} > 0\right\} \cup \mathcal{L}_{>0}(A)$. ☐

Proof of Equation (49). We start with the first part of Equation (49). Due to Equation (17) and Proposition 19, it suffices to prove that, for any $p \in \mathcal{V}(A)$, $\min_{x \in A} p(\boldsymbol{\theta}^\circ_x) > 0 \Rightarrow p \in \mathcal{H}_{\mathrm{SC},A,\mathbf{0}}$ and $\min_{x \in A} p(\boldsymbol{\theta}^\circ_x) < 0 \Rightarrow p \notin \mathcal{H}_{\mathrm{SC},A,\mathbf{0}}$.

First, assume that $\min_{x\in A} p(\boldsymbol{\theta}_x^\circ) < 0$. Then there is some $y \in A$ for which $p(\boldsymbol{\theta}_y^\circ) < 0$. Hence, since $p_{|\Sigma_{\{y\}}} = p(\boldsymbol{\theta}_y^\circ) < 0$, we find that $\{y\} \in \min S_{A,\mathbf{0}}(p)$ and therefore also that $p \notin \mathcal{H}_{\mathrm{SC},A,\mathbf{0}}$, by Proposition 18.

Next, assume that $\min_{x\in A} p(\boldsymbol{\theta}_x^\circ) > 0$. Then $p_{|\Sigma_{\{x\}}} = p(\boldsymbol{\theta}_x^\circ) > 0$ for all $x \in A$, implying that $\min S_{A,\mathbf{0}}(p) = \{\{x\}\colon x \in A\}$ and therefore also, since $p \neq 0$, that $p \in \mathcal{H}_{\mathrm{SC},A,\mathbf{0}}$, by Proposition 18.

We now turn to the second part of Equation (49). Due to Equation (17) and Proposition 19, it suffices to prove that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and any $p \in \mathcal{V}(A)$, $\min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} p(\boldsymbol{\theta}) > 0 \Rightarrow p \in \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}$ and $\min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} p(\boldsymbol{\theta}) < 0 \Rightarrow p \notin \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}$.

First, assume that $\min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} p(\boldsymbol{\theta}) < 0$. Then there is some $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_{A[\check{\boldsymbol{m}}]})$ for which $p(\boldsymbol{\theta}) < 0$, implying that $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \neq 0$ and $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \notin \mathcal{V}^{++}(A[\check{\boldsymbol{m}}])$. Hence, we find that $A[\check{\boldsymbol{m}}] \in \min S_{A,\check{\boldsymbol{m}}}(p)$ and therefore also that $p \notin \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}$, by Proposition 18.

Next, assume that $\min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} p(\boldsymbol{\theta}) > 0$. Then $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \neq 0$ and $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \in \mathcal{V}^{++}(A[\check{\boldsymbol{m}}])$. Hence, we find that $\min S_{A,\check{\boldsymbol{m}}}(p) = \{A[\check{\boldsymbol{m}}]\}$ and therefore also, since $p \neq 0$, that $p \in \mathcal{H}_{\mathrm{SC},A,\check{\boldsymbol{m}}}$, by Proposition 18. □

Proof of Equation (52). The first part of Equation (52) is a trivial consequence of Equation (51). For the second part, consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and any $f \in \mathcal{L}(A)$. Then, combining Equations (50) and (30):

$$\underline{P}^1_{\mathrm{SC},A}(f|\check{\boldsymbol{m}}) = \min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} \sum_{x\in A} f(x)\theta_x = \min_{\boldsymbol{\theta}\in\Sigma_{A[\check{\boldsymbol{m}}]}} \sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\theta_x = \min_{x\in A[\check{\boldsymbol{m}}]} f(x). \qquad \square$$

Lemma 45. *Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\emptyset \neq K \subseteq A$. Then $(p \circ R_\rho)_{|\Sigma_K} \neq 0 \Leftrightarrow p_{|\Sigma_{\rho(K)}} \neq 0$.*

Proof of Lemma 45. First, assume that $p_{|\Sigma_{\rho(K)}} \neq 0$, so there is some $\boldsymbol{\vartheta} \in \Sigma_{\rho(K)}$ such that $p(\boldsymbol{\vartheta}) \neq 0$. Now choose any $\boldsymbol{\theta} \in \Sigma_K$ such that $R_\rho(\boldsymbol{\theta}) = \boldsymbol{\vartheta}$. Then, clearly, $(p \circ R_\rho)(\boldsymbol{\theta}) = p(R_\rho(\boldsymbol{\theta})) = p(\boldsymbol{\vartheta}) \neq 0$ and therefore $(p \circ R_\rho)_{|\Sigma_K} \neq 0$.

Assume, conversely, that $(p\circ R_\rho)_{|\Sigma_K} \neq 0$, so there is some $\boldsymbol{\theta} \in \Sigma_K$ such that $(p\circ R_\rho)(\boldsymbol{\theta}) \neq 0$. If we let $\boldsymbol{\vartheta} := R_\rho(\boldsymbol{\theta})$, then $\boldsymbol{\vartheta} \in \Sigma_{\rho(K)}$ and $p(\boldsymbol{\vartheta}) = p(R_\rho(\boldsymbol{\theta})) = (p \circ R_\rho)(\boldsymbol{\theta}) \neq 0$. Hence, $p_{|\Sigma_{\rho(K)}} \neq 0$. □

Lemma 46. *Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$, and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$. Then*

$$S_{A,\boldsymbol{m}}(p \circ R_\rho) = \left\{K \subseteq A\colon A[\boldsymbol{m}] \subseteq K \text{ and } \rho(K) \in S_{D,R_\rho(\boldsymbol{m})}(p)\right\},$$

and therefore

$$\rho(S_{A,\boldsymbol{m}}(p \circ R_\rho)) = S_{D,R_\rho(\boldsymbol{m})}(p) \text{ and } \rho(\min S_{A,\boldsymbol{m}}(p \circ R_\rho)) = \min S_{D,R_\rho(\boldsymbol{m})}(p).$$

Proof of Lemma 46. We start by proving the first statement. First, assume that $K \in S_{A,\boldsymbol{m}}(p\circ R_\rho)$, implying that $\emptyset \neq K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $(p \circ R_\rho)_{|\Sigma_K} \neq 0$. Then $\emptyset \neq \rho(K) \subseteq D$, $D[R_\rho(\boldsymbol{m})] = \rho(A[\boldsymbol{m}]) \subseteq \rho(K)$ and, by Lemma 45, $p_{|\Sigma_{\rho(K)}} \neq 0$. Hence, $\rho(K) \in S_{D,R_\rho(\boldsymbol{m})}(p)$. Conversely, assume that $K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $\rho(K) \in S_{D,R_\rho(\boldsymbol{m})}(p)$. Then $\emptyset \neq \rho(K)$, which implies that $\emptyset \neq K$, and also $p_{|\Sigma_{\rho(K)}} \neq 0$, which, by Lemma 45, implies that $(p \circ R_\rho)_{|\Sigma_K} \neq 0$. Hence, $K \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$.

The first statement implies that $\rho(S_{A,\boldsymbol{m}}(p \circ R_\rho)) \subseteq S_{D,R_\rho(\boldsymbol{m})}(p)$ and therefore, in order to prove the second statement, it suffices to show that $S_{D,R_\rho(\boldsymbol{m})}(p) \subseteq \rho(S_{A,\boldsymbol{m}}(p \circ R_\rho))$ or, equivalently, that for every $L \in S_{D,R_\rho(\boldsymbol{m})}(p)$, there is some $K \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$ such that $\rho(K) = L$. So choose any $L \in S_{D,R_\rho(\boldsymbol{m})}(p)$ and let $K := \{x \in A \colon \rho(x) \in L\} = \rho^{-1}(L)$. Then $\rho(K) = L$ because ρ is onto, and since $\rho(A[\boldsymbol{m}]) = D[R_\rho(\boldsymbol{m})] \subseteq L$, it follows that $A[\boldsymbol{m}] \subseteq K$. Hence, by the first statement, $K \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$.

To prove the third statement, first assume that $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$, implying that $K \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$ and that, for all $K' \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$, $K' \not\subset K$. By the second statement, $\rho(K) \in S_{D,R_\rho(\boldsymbol{m})}(p)$. To prove that $\rho(K) \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$, assume *ex absurdo* that there is some $L \in S_{D,R_\rho(\boldsymbol{m})}(p)$ such that $L \subset \rho(K)$. Let $K' := \{x \in K \colon \rho(x) \in L\} = K \cap \rho^{-1}(L)$. Then $K' \subset K$ and $\rho(K') = L$, and therefore, by Lemma 45, $(p \circ R_\rho)_{|\Sigma_{K'}} \neq 0$, because $K' \neq \emptyset$ and $p_{|\Sigma_L} \neq 0$. Since $L \in S_{D,R_\rho(\boldsymbol{m})}(p)$, we see that $\rho(A[\boldsymbol{m}]) = D[R_\rho(\boldsymbol{m})] \subseteq L$ and therefore $A[\boldsymbol{m}] \subseteq \rho^{-1}(L)$. Since $K \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$, we also know that $A[\boldsymbol{m}] \subseteq K$, and therefore $A[\boldsymbol{m}] \subseteq K \cap \rho^{-1}(L) = K'$. This tells us that $K' \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$, a contradiction. Assume, conversely, that $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$, implying that $L \in S_{D,R_\rho(\boldsymbol{m})}(p)$. Then, by the second statement, there is some $K' \in S_{A,\boldsymbol{m}}(p \circ R_\rho)$ such that $\rho(K') = L$. Hence, there is some $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$ such that $K \subseteq K'$ and therefore $\rho(K) \subseteq \rho(K') = L$. Since $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$ and since, due to the second statement, $\rho(K) \in S_{D,R_\rho(\boldsymbol{m})}(p)$, we also have that $\rho(K) \not\subset L$ and therefore $\rho(K) = L$. ∎

Lemma 47. *Let $\emptyset \neq K \subseteq A$ and let p be any polynomial on Σ_A. Then for any $n \geq \deg(p)$: $b^n_{p_{|\Sigma_K}} = b^n_{p|\mathcal{N}^n_K}$.*

Proof of Lemma 47. It follows from

$$p(\theta) = \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} b^n_p(\boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}}(\theta) \text{ for all } \theta \in \Sigma_A$$

that for all $\vartheta \in \Sigma_K$:

$$\begin{aligned} p_{|\Sigma_K}(\vartheta) &= \sum_{\boldsymbol{m} \in \mathcal{N}^n_A} b^n_p(\boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}}(i_A(\vartheta)) = \sum_{\boldsymbol{m} \in \mathcal{N}^n_A \colon A[\boldsymbol{m}] \subseteq K} b^n_p(\boldsymbol{m}) \mathrm{B}_{A,\boldsymbol{m}}(i_A(\vartheta)) \\ &= \sum_{\boldsymbol{n} \in \mathcal{N}^n_K} b^n_{p|\mathcal{N}^n_K}(\boldsymbol{n}) \mathrm{B}_{K,\boldsymbol{n}}(\vartheta), \end{aligned}$$

and this completes the proof. ∎

Lemma 48. *Consider any category sets A and D such that there is an onto map $\rho \colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\emptyset \neq K \subseteq A$. Then:*

(i) $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{V}^+(K) \Leftrightarrow p_{|\Sigma_{\rho(K)}} \in \mathcal{V}^+(\rho(K))$;

(ii) $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{V}^{++}(K) \Leftrightarrow p_{|\Sigma_{\rho(K)}} \in \mathcal{V}^{++}(\rho(K))$.

Proof of Lemma 48. The first statement follows from the fact that, for all $n \geq \deg(p)$:

$$b^n_{(p\circ R_\rho)_{|\Sigma_K}} > 0 \Leftrightarrow b^n_{(p\circ R_\rho)|\mathcal{N}^n_K} > 0 \Leftrightarrow (b^n_p \circ R_\rho)_{|\mathcal{N}^n_K} > 0 \Leftrightarrow b^n_{p|\mathcal{N}^n_{\rho(K)}} > 0 \Leftrightarrow b^n_{p_{|\Sigma_{\rho(K)}}} > 0,$$

where the first and last equivalence are due to Lemma 47, the second equivalence follows from Proposition 39, and the third equivalence holds because $R_\rho(\mathcal{N}^n_K) = \mathcal{N}^n_{\rho(K)}$.

We now turn to the second statement, where we have to prove that the following statements are equivalent:

(a) $(\forall \boldsymbol{\theta} \in \operatorname{int}(\Sigma_K))p(R_\rho(\boldsymbol{\theta})) > 0$;

(b) $(\forall \boldsymbol{\vartheta} \in \operatorname{int}(\Sigma_{\rho(K)}))p(\boldsymbol{\vartheta}) > 0$.

First assume that (a) holds, and consider any $\boldsymbol{\vartheta} \in \operatorname{int}(\Sigma_{\rho(K)})$. We have to prove that $p(\boldsymbol{\vartheta}) > 0$. We construct a $\boldsymbol{\theta} \in \Sigma_K$ as follows. Consider any $z \in \rho(K)$. For all $x \in \rho^{-1}(\{z\}) \cap K$, choose the $\theta_x > 0$ in such a way that $\sum_{x\in K\colon \rho(x)=z} \theta_x = \vartheta_z$. In this way, we have found a $\boldsymbol{\theta} \in \Sigma_K$ satisfying $R_\rho(\boldsymbol{\theta}) = \boldsymbol{\vartheta}$, and such that moreover $\theta_x > 0$ for all $x \in K$, whence $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$. We now infer from (a) that indeed $p(\boldsymbol{\vartheta}) = p(R_\rho(\boldsymbol{\theta})) > 0$.

Assume, conversely, that (b) holds, and consider any $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$. Then, for any $z \in D$, $R_\rho(\boldsymbol{\theta})_z > 0$ if $z \in \rho(K)$ and $R_\rho(\boldsymbol{\theta})_z = 0$ otherwise. This means that $R_\rho(\boldsymbol{\theta}) \in \operatorname{int}(\Sigma_{\rho(K)})$ and we infer from (b) that indeed $p(R_\rho(\boldsymbol{\theta})) > 0$. □

Proposition 49. *Φ_{SC} is representation insensitive.*

Proof of Proposition 49. We use the characterisation of representation insensitivity in Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then, by Proposition 19, we need to prove that $p \circ R_\rho \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}} \Leftrightarrow p \in \mathcal{H}_{\mathrm{SC},D,R_\rho(\boldsymbol{m})}$.

First, assume that $p \in \mathcal{H}_{\mathrm{SC},D,R_\rho(\boldsymbol{m})}$, which, by Proposition 18, implies that $p \neq 0$ and that $p_{|\Sigma_L} \in \mathcal{V}^{++}(L)$ for all $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$. Applying Lemma 45 with $K = A$, we infer from $p \neq 0$ that $p \circ R_\rho \neq 0$. Consider now any $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$. Then, by Lemma 46, $\rho(K) \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$, implying that, due to the assumption, $p_{|\Sigma_{\rho(K)}} \in \mathcal{V}^{++}(\rho(K))$. Since $K \neq \emptyset$, we can apply Lemma 48 to find that $(p\circ R_\rho)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$. Hence, by Proposition 18, $p \circ R_\rho \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$.

Assume, conversely, that $p\circ R_\rho \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$, which, by Proposition 18, implies that $p\circ R_\rho \neq 0$ and that $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ for all $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$. Applying Lemma 45, with $K = A$, we infer from $p \circ R_\rho \neq 0$ that $p \neq 0$. Now, consider any $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$, then by Lemma 46, there is some $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$ such that $\rho(K) = L$. Since $K \neq \emptyset$ and, by assumption, $(p\circ R_\rho)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$, we infer from Lemma 48 that $p_{|\Sigma_L} \in \mathcal{V}^{++}(L)$. Hence, by Proposition 18, $p \in \mathcal{H}_{\mathrm{SC},D,R_\rho(\boldsymbol{m})}$. □

Lemma 50. *Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $K \subseteq A$ such that $K \cap B \neq \emptyset$ and any $r \in \mathbb{N}_0$. Then $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0 \Leftrightarrow p_{|\Sigma_{K\cap B}} \neq 0$.*

Proof of Lemma 50. We may assume without loss of generality that $r + \deg(p) > 0$, as the proof is trivial otherwise.

First, assume that $p_{|\Sigma_{K\cap B}} \neq 0$, which means that there is some $\boldsymbol{\vartheta} \in \Sigma_{K\cap B}$ such that $p(\boldsymbol{\vartheta}) \neq 0$. Then $\boldsymbol{\theta} := i_A(\boldsymbol{\vartheta}) \in \Sigma_K$, and we infer from Proposition 9 that $\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) = p(\boldsymbol{\vartheta}) \neq 0$ and therefore $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0$.

Assume, conversely, that $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0$, which means, due to the continuity of polynomials, that there is some $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$ such that $\mathrm{I}^r_{B,A}(p|\boldsymbol{\theta}) \neq 0$. We now infer from $K \cap B \neq \emptyset$ that $\theta_B > 0$, so Proposition 10 guarantees that $p(\boldsymbol{\theta}|^+_B) \neq 0$. Since $\boldsymbol{\theta}|^+_B \in \Sigma_{K\cap B}$, we find that $p_{|\Sigma_{K\cap B}} \neq 0$. □

Lemma 51. *Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$ any $r \in \mathbb{N}_0$ such that $r + \deg(p) > 0$, and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then*

$$S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p)) = \left\{K \subseteq A \colon A[\boldsymbol{m}] \subseteq K \text{ and } K \cap B \in S_{B,r_B(\boldsymbol{m})}(p)\right\},$$

and therefore

$$S_{B,r_B(\boldsymbol{m})}(p) = \left\{K \cap B \colon K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))\right\}$$

and

$$\min S_{B,r_B(\boldsymbol{m})}(p) = \left\{K \cap B \colon K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))\right\}.$$

Proof of Lemma 51. We begin with the first statement. First, assume that $K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$ and therefore that $\emptyset \neq K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0$. Then $K \subseteq A$ implies that $K \cap B \subseteq B$, $A[\boldsymbol{m}] \subseteq K$ implies that $B[r_B(\boldsymbol{m})] = A[\boldsymbol{m}] \cap B \subseteq K \cap B$. Moreover, $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0$ together with Proposition 10 and $r + \deg(p) > 0$ implies that $K \cap B \neq \emptyset$, which in turn, by Lemma 50, implies that $p_{|\Sigma_{K\cap B}} \neq 0$. Hence, $K \cap B \in S_{B,r_B(\boldsymbol{m})}(p)$. Conversely, assume that $K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $K \cap B \in S_{B,r_B(\boldsymbol{m})}(p)$. Then $K \cap B \neq \emptyset$, implying that $K \neq \emptyset$, and $p_{|\Sigma_{K\cap B}} \neq 0$, which, by Lemma 50, implies that $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \neq 0$. Hence, $K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$.

In order to prove the second statement, it clearly suffices to show that $S_{B,r_B(\boldsymbol{m})}(p) \subseteq \{K \cap B \colon K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))\}$, since the converse inclusion follows directly from the first statement. So consider any $L \in S_{B,r_B(\boldsymbol{m})}(p)$ and let $K \coloneqq L \cup A[\boldsymbol{m}]$. Then $K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and $K \cap B = L \cup (A[\boldsymbol{m}] \cap B) = L \cup B[r_B(\boldsymbol{m})] = L$. Hence, by the first statement, indeed $K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$.

To prove the third statement, first assume that $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$, implying that in particular $K \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$. Then, by the second statement, $K \cap B \in S_{B,r_B(\boldsymbol{m})}(p)$. To prove that $K \cap B \in \min S_{B,r_B(\boldsymbol{m})}(p)$, consider any $L \in S_{B,r_B(\boldsymbol{m})}(p)$ such that $L \subseteq K \cap B$, and let $K' \coloneqq L \cup A[\boldsymbol{m}]$. Then, by an argument identical to the one used in the proof of the second statement, $K' \cap B = L$ and $K' \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$. However, since $K' \cap B = L \subseteq K \cap B$ and $K' \setminus B = A[\boldsymbol{m}] \setminus B \subseteq K \setminus B$, we find that $K' = (K' \cap B) \cup (K' \setminus B) \subseteq (K \cap B) \cup (K \setminus B) = K$, and therefore $K' = K$, by assumption. Hence indeed $L = K' \cap B = K \cap B$. Assume, conversely, that $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$, implying that $L \in S_{B,r_B(\boldsymbol{m})}(p)$. Then, by the second statement, there is some $K' \in S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$ such that $K' \cap B = L$, so there is some $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$ such that $K \subseteq K'$ and therefore $K \cap B \subseteq K' \cap B = L$. Since $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$ and, by the second statement, $K \cap B \in S_{B,r_B(\boldsymbol{m})}(p)$, we also have that $K \cap B = L$. □

Lemma 52. *Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $K \subseteq A$ such that $K \cap B \neq \emptyset$ and any $r \in \mathbb{N}_0$. Then $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \in \mathcal{V}^{++}(K) \Leftrightarrow p_{|\Sigma_{K\cap B}} \in \mathcal{V}^{++}(K \cap B)$.*

Proof of Lemma 52. We may assume without loss of generality that $r + \deg(p) > 0$, as the proof is trivial otherwise. Using Proposition 10, and considering that, since $K \cap B \neq \emptyset$, $\theta_B > 0$ for any $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$, it then suffices to prove that the following statements are equivalent:

(a) $(\forall \boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)) p(\boldsymbol{\theta}|_B^+) > 0$;

(b) $(\forall \boldsymbol{\vartheta} \in \operatorname{int}(\Sigma_{K\cap B})) p(\boldsymbol{\vartheta}) > 0$.

First assume that (a) holds, and consider any $\boldsymbol{\vartheta} \in \operatorname{int}(\Sigma_{K\cap B})$. We have to prove that $p(\boldsymbol{\vartheta}) > 0$. We construct a $\boldsymbol{\theta} \in \Sigma_K$ as follows. For any $x \in K \setminus B$, choose $\theta_x > 0$ in such a way that $\kappa := \sum_{x\in K\setminus B} \theta_x < 1$, which is always possible. And for any $x \in K \cap B$, let $\theta_x := (1-\kappa)\vartheta_x > 0$. Then it follows from this construction that $\theta_B = 1 - \kappa > 0$, $\boldsymbol{\theta}|_B^+ = \boldsymbol{\vartheta}$ and $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$, so we infer from (a) that indeed $p(\boldsymbol{\vartheta}) = p(\boldsymbol{\theta}|_B^+) > 0$.

Assume, conversely, that (b) holds, and consider any $\boldsymbol{\theta} \in \operatorname{int}(\Sigma_K)$. Then $\theta_B > 0$ because $K \cap B \neq 0$ and therefore, for all $z \in B$, $(\boldsymbol{\theta}|_B^+)_z > 0 \Leftrightarrow z \in K \cap B$. Hence $\boldsymbol{\theta}|_B^+ \in \operatorname{int}(\Sigma_{K\cap B})$, so we infer from (b) that $p(\boldsymbol{\theta}|_B^+) > 0$. □

Proposition 53. Φ_{SC} *is specific.*

Proof of Proposition 53. We use the characterisation of specificity in Theorem 11. Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, and any $r \in \mathbb{N}_0$. Then, by Proposition 19, we need to prove that $\mathrm{I}_{B,A}^r(p) \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}} \Leftrightarrow p \in \mathcal{H}_{\mathrm{SC},B,r_B(\boldsymbol{m})}$.

First, assume that $p \in \mathcal{H}_{\mathrm{SC},B,r_B(\boldsymbol{m})}$, which, by Proposition 18, implies that $p \neq 0$ and that $p_{|\Sigma_L} \in \mathcal{V}^{++}(L)$ for all $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$. Applying Lemma 50 with $K = A$, we infer from $p \neq 0$ that $\mathrm{I}_{B,A}^r(p) \neq 0$. Consider any $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}_{B,A}^r(p))$, then by Lemma 51, $K \cap B \in \min S_{B,r_B(\boldsymbol{m})}(p)$, implying that, due to the assumption, $p_{|\Sigma_{K\cap B}} \in \mathcal{V}^{++}(K \cap B)$. Since $K \cap B \neq 0$, we can apply Lemma 52 to find that $\mathrm{I}_{B,A}^r(p)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$. Hence, again by Proposition 18, $\mathrm{I}_{B,A}^r(p) \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$.

Assume, conversely, that $\mathrm{I}_{B,A}^r(p) \in \mathcal{H}_{\mathrm{SC},A,\boldsymbol{m}}$, which, by Proposition 18, implies that $\mathrm{I}_{B,A}^r(p) \neq 0$ and that $\mathrm{I}_{B,A}^r(p)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$ for all $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}_{B,A}^r(p))$. From Lemma 50 with $K = A$, and from $\mathrm{I}_{B,A}^r(p) \neq 0$, we infer that $p \neq 0$. Consider any $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$, then, by Lemma 51, there is some $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}_{B,A}^r(p))$ such that $K \cap B = L$. Since therefore $K \cap B \neq 0$ and since, by assumption, $\mathrm{I}_{B,A}^r(p)_{|\Sigma_K} \in \mathcal{V}^{++}(K)$, we infer from Lemma 52 that $p_{|\Sigma_L} \in \mathcal{V}^{++}(L)$. Hence, by Proposition 18, $p \in \mathcal{H}_{\mathrm{SC},B,r_B(\boldsymbol{m})}$. □

Proof of Theorem 20. This is an immediate consequence of Propositions 17 [coherence], 49 [representation insensitivity] and 53 [specificity]. □

E.7 Proofs of Results in Section 12

Proof of Equation (54). Consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $p \in \mathcal{V}(A)$. Then

$$\begin{aligned} p \in \mathcal{H}_{\mathrm{IDM},A}^s \rfloor \check{\boldsymbol{m}} \Leftrightarrow \mathrm{B}_{A,\boldsymbol{m}} p \in \mathcal{H}_{\mathrm{IDM},A}^s &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_A^s)\, \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}} p | \boldsymbol{\alpha}) > 0 \\ &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_A^s)\, \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}} | \boldsymbol{\alpha})\, \mathrm{Di}_A(p | \check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0 \\ &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_A^s)\, \mathrm{Di}_A(p | \check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0, \end{aligned}$$

where the third equivalence follows from the Updating Property of the Dirichlet expectation [Proposition 31]. □

Proof of Equation (59). Consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then, combining Equations (56) and (58) for $\hat{n} = 1$:

$$\mathcal{D}_{\mathrm{IDM},A}^{s,1} \rfloor \check{\boldsymbol{m}} = \Big\{ f \in \mathcal{L}(A) \colon (\forall \boldsymbol{\alpha} \in \Delta_A^s) \sum_{x\in A} f(x) \frac{\check{m}_x + \alpha_x}{\check{m}_A + \alpha_A} > 0 \Big\}.$$

Now consider any $f \in \mathcal{L}(A)$. Then for all $\boldsymbol{\alpha} \in \Delta_A^s$:

$$\sum_{x\in A} f(x)\frac{\check{m}_x+\alpha_x}{\check{m}_A+\alpha_A} > 0 \Leftrightarrow \sum_{x\in A} f(x)(\check{m}_x+\alpha_x) > 0 \Leftrightarrow \sum_{x\in A} f(x)\frac{\alpha_x}{s} > -\frac{1}{s}\sum_{x\in A} f(x)\check{m}_x.$$

Combining the equations above, and letting $c := -\frac{1}{s}\sum_{x\in A} f(x)\check{m}_x$ for ease of notation, we find that:

$$f \in \mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}} \Leftrightarrow (\forall s' \in (0,s))(\forall \boldsymbol{t} \in \operatorname{int}(\Sigma_A))\frac{s'}{s}\sum_{x\in A} f(x)t_x > c. \tag{95}$$

If $f \not> c$, then there is some $y \in A$ for which $f(y) < c$ and therefore, by Statement (95), $f \notin \mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}$ [choose s' and t_y close enough to s and 1, respectively]. If $f = c$, then due to the definition of c, $f = c = 0$. Hence, again by Statement (95), $f \notin \mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}$. Finally, let us see what happens if $f > c$. Then clearly $c \leq 0$. Consider any $s' \in (0,s)$ and any $\boldsymbol{t} \in \operatorname{int}(\Sigma_A)$. Then since $f > c$, $\sum_{x\in A} f(x)t_x > c$ and therefore also, since $c \leq 0$, $\frac{s'}{s}\sum_{x\in A} f(x)t_x > \frac{s'}{s}c \geq c$. Hence $f \in \mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}$ by Statement (95). $\square$

Proof of Equation (60). Consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ and any $f \in \mathcal{L}(A)$. Then by combining Equations (57) and (58):

$$\begin{aligned}
\underline{P}^{s,1}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}}) &= \inf_{\boldsymbol{\alpha}\in\Delta_A^s}\sum_{x\in A} f(x)\frac{\check{m}_x+\alpha_x}{\check{m}_A+\alpha_A} = \inf_{s'\in(0,s)}\inf_{\boldsymbol{t}\in\operatorname{int}(\Sigma_A)}\sum_{x\in A} f(x)\frac{\check{m}_x+s't_x}{\check{m}_A+s'}\\
&= \inf_{s'\in(0,s)}\left(\frac{1}{\check{m}_A+s'}\sum_{x\in A} f(x)\check{m}_x + \frac{s'}{\check{m}_A+s'}\inf_{\boldsymbol{t}\in\operatorname{int}(\Sigma_A)}\sum_{x\in A} f(x)t_x\right)\\
&= \inf_{s'\in(0,s)}\left(\frac{1}{\check{m}_A+s'}\sum_{x\in A} f(x)\check{m}_x + \frac{s'}{\check{m}_A+s'}\min f\right)\\
&= \frac{1}{\check{m}_A+s}\sum_{x\in A} f(x)\check{m}_x + \frac{s}{\check{m}_A+s}\min f,
\end{aligned}$$

where the last equality follows from $\min f \leq \sum_{x\in A} f(x)\frac{\check{m}_x}{\check{m}_A}$, a property of convex combinations. $\square$

Proof of Theorem 21. For coherence, if we fix any category set A, then we must prove that $\mathcal{H}^s_{\mathrm{IDM},A}$ satisfies the requirements B1–B3 of Bernstein coherence. This is trivial from the definition of $\mathcal{H}^s_{\mathrm{IDM},A}$, the linearity of the Dirichlet expectation operator, and the fact that the Dirichlet expectation of any Bernstein basis polynomial is positive.

Next, we turn to representation insensitivity, and use its characterisation in Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$. Then, using the Pooling Property [Proposition 33] of the Dirichlet expectation and Equation (54), we find that indeed:

$$\begin{aligned}
(p\circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{IDM},A} &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_A^s)\,\mathrm{Di}_A(p\circ R_\rho|\boldsymbol{m}+\boldsymbol{\alpha}) > 0\\
&\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta_A^s)\,\mathrm{Di}_D(p|R_\rho(\boldsymbol{m}+\boldsymbol{\alpha})) > 0\\
&\Leftrightarrow (\forall \boldsymbol{\beta} \in \Delta_D^s)\,\mathrm{Di}_D(p|R_\rho(\boldsymbol{m})+\boldsymbol{\beta}) > 0 \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},D},
\end{aligned}$$

where the third equivalence follows from the equality $\Delta^s_D = R_\rho(\Delta^s_A)$.

Finally, we turn to specificity, and use its characterisation in Theorem 11. Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$ and any $r \in \mathbb{N}_0$. Then, using the Restriction Property [Proposition 34] of the Dirichlet expectation and Equation (54), we find that indeed:

$$\begin{aligned}\mathrm{I}^r_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{IDM},A} &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta^s_A)\, \mathrm{Di}_A(\mathrm{I}^r_{B,A}(p)|\boldsymbol{m}+\boldsymbol{\alpha}) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\alpha} \in \Delta^s_A)\, \mathrm{Di}_B(p|r_B(\boldsymbol{m}+\boldsymbol{\alpha})) > 0\\ &\Leftrightarrow (\forall \boldsymbol{\beta} \in \Delta^s_B)\, \mathrm{Di}_B(p|r_B(\boldsymbol{m})+\boldsymbol{\beta}) > 0 \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},B},\end{aligned}$$

where the third equivalence follows from $\Delta^s_B = r_B(\Delta^s_A)$. ☐

E.8 Proofs of Results in Section 13

Lemma 54. *For any $p_1, p_2 \in \mathcal{H}^s_{\mathrm{SI},A}$: $S_{A,\mathbf{0}}(p_1+p_2) = S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2)$.*

Proof of Lemma 54. First, consider any $K \in S_{A,\mathbf{0}}(p_1+p_2)$, meaning that $\emptyset \neq K \subseteq A$ and $(p_1+p_2)_{|\Sigma_K} \neq 0$. Assume, *ex absurdo*, that $K \notin S_{A,\mathbf{0}}(p_1)$ and $K \notin S_{A,\mathbf{0}}(p_2)$. Then $p_{1|\Sigma_K} = 0$ and $p_{2|\Sigma_K} = 0$ and therefore $(p_1+p_2)_{|\Sigma_K} = 0$, which is a contradiction. Hence indeed $K \in S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2)$.

Next, consider any $K \in S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2)$, implying that $\emptyset \neq K \subseteq A$. Then there is at least one $K' \in \min(S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2))$ such that $K' \subseteq K$, and we can assume without loss of generality that $K' \in S_{A,\mathbf{0}}(p_1)$. Since $K' \in \min(S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2))$, we have that $L \not\subset K'$ for all $L \in S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2)$, and therefore $K' \in \min S_{A,\mathbf{0}}(p_1)$. This already tells us that $p_{1|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$. There are now two possibilities. The first one is that $K' \in S_{A,\mathbf{0}}(p_2)$, and then, in very much the same way as as above, we find that $p_{2|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$. Hence, due to the Bernstein coherence [B3] of $\mathcal{H}^s_{\mathrm{IDM},K'}$, $(p_1+p_2)_{|\Sigma_{K'}} = p_{1|\Sigma_{K'}} + p_{2|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$. The second possibility is that $K' \notin S_{A,\mathbf{0}}(p_2)$, and then $p_{2|\Sigma_{K'}} = 0$ since $K' \neq \emptyset$, so we find, here too, that $(p_1+p_2)_{|\Sigma_{K'}} = p_{1|\Sigma_{K'}} + p_{2|\Sigma_{K'}} = p_{1|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$. In both cases, therefore, $(p_1+p_2)_{|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$, and the Bernstein coherence [B1] of $\mathcal{H}^s_{\mathrm{IDM},K'}$ allows us to conclude that $(p_1+p_2)_{|\Sigma_{K'}} \neq 0$. Since $K' \subseteq K$, we find that also $(p_1+p_2)_{|\Sigma_K} \neq 0$ and therefore that $K \in S_{A,\mathbf{0}}(p_1+p_2)$. ☐

Proof of Proposition 22. Since $0 \notin \mathcal{H}^s_{\mathrm{SI},A}$, we are left to prove that $\mathcal{V}^+(A) \subseteq \mathcal{H}^s_{\mathrm{SI},A}$ and that, for all $\lambda > 0$ and $p, p_1, p_2 \in \mathcal{H}^s_{\mathrm{SI},A}$, $\lambda p \in \mathcal{H}^s_{\mathrm{SI},A}$ and $p_1+p_2 \in \mathcal{H}^s_{\mathrm{SI},A}$.

First, consider any $\lambda > 0$ and $p \in \mathcal{H}^s_{\mathrm{SI},A}$. Then, clearly, $S_{A,\mathbf{0}}(\lambda p) = S_{A,\mathbf{0}}(p)$ and therefore $\min S_{A,\mathbf{0}}(\lambda p) = \min S_{A,\mathbf{0}}(p)$. Then for any $K \in \min S_{A,\mathbf{0}}(\lambda p)$, we have that $K \in \min S_{A,\mathbf{0}}(p)$, which, since $p \in \mathcal{H}^s_{\mathrm{SI},A}$, implies that $p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$ and therefore, due to the Bernstein coherence of $\mathcal{H}^s_{\mathrm{IDM},K}$, that $(\lambda p)_{|\Sigma_K} = \lambda(p_{|\Sigma_K}) \in \mathcal{H}^s_{\mathrm{IDM},K}$. Furthermore, since $p \neq 0$ also $\lambda p \neq 0$, and therefore $\lambda p \in \mathcal{H}^s_{\mathrm{SI},A}$.

Next, consider any $p_1, p_2 \in \mathcal{H}^s_{\mathrm{SI},A}$. Then $p_1 \neq 0$ and $p_2 \neq 0$, implying that $S_{A,\mathbf{0}}(p_1) \neq \emptyset$ and $S_{A,\mathbf{0}}(p_2) \neq \emptyset$, and therefore $S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2) \neq \emptyset$. Applying Lemma 54, we find that $S_{A,\mathbf{0}}(p_1+p_2) \neq \emptyset$, so there is some K such that $\emptyset \neq K \subseteq A$, $(p_1+p_2)_{|\Sigma_K} \neq 0$ and therefore $p_1+p_2 \neq 0$. Then for any $K' \in \min S_{A,\mathbf{0}}(p_1+p_2)$, or equivalently, due to Lemma 54, $K' \in \min(S_{A,\mathbf{0}}(p_1) \cup S_{A,\mathbf{0}}(p_2))$. Then, by applying the same reasoning as in the second part of the proof of Lemma 54, we find that $(p_1+p_2)_{|\Sigma_{K'}} \in \mathcal{H}^s_{\mathrm{IDM},K'}$. Hence, $p_1+p_2 \in \mathcal{H}^s_{\mathrm{SI},A}$.

Since we have already shown that $\mathcal{H}^s_{\mathrm{SI},A}$ is closed under taking positive linear combinations, and since $\mathcal{V}^+(A)$ consists of positive linear combinations of Bernstein basis polynomials, we only need to show that $\mathcal{H}^s_{\mathrm{SI},A}$ contains all Bernstein basis polynomials in order to prove that $\mathcal{V}^+(A) \subseteq \mathcal{H}^s_{\mathrm{SI},A}$. So consider any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then, for any K such that $\emptyset \neq K \subseteq A$, we have that $\mathrm{B}_{A,\boldsymbol{m}|\Sigma_K} = \mathrm{B}_{K,r_K(\boldsymbol{m})}$ if $A[\boldsymbol{m}] \subseteq K$, and $\mathrm{B}_{A,\boldsymbol{m}|\Sigma_K} = 0$ otherwise. This implies that $S_{A,\boldsymbol{0}}(\mathrm{B}_{A,\boldsymbol{m}}) = \{\emptyset \neq K \subseteq A \colon A[\boldsymbol{m}] \subseteq K\}$ and that, due to the Bernstein coherence of $\mathcal{H}^s_{\mathrm{IDM},K}$, $\mathrm{B}_{A,\boldsymbol{m}|\Sigma_K} = \mathrm{B}_{K,r_K(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},K}$ for all $K \in S_{A,\boldsymbol{0}}(\mathrm{B}_{A,\boldsymbol{m}})$. Hence, $\mathrm{B}_{A,\boldsymbol{m}|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$ for all $K \in \min S_{A,\boldsymbol{0}}(\mathrm{B}_{A,\boldsymbol{m}})$. Since also $\mathrm{B}_{A,\boldsymbol{m}} \neq 0$, we find that indeed $\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{SI},A}$. $\square$

Proof of Proposition 23. We first prove that $\mathcal{H}^s_{\mathrm{SI},A}\rfloor\boldsymbol{m} \subseteq \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$. Consider any $p \in \mathcal{V}(A)$ such that $p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{SI},A}$, meaning that $p\mathrm{B}_{A,\boldsymbol{m}} \neq 0$ and that $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$ for all $K \in \min S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$. We set out to prove that $p \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$. Since, clearly, $p \neq 0$, it suffices to show that $p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ for all $K \in \min S_{A,\boldsymbol{m}}(p)$. So consider any $K \in \min S_{A,\boldsymbol{m}}(p)$, implying that $A[\boldsymbol{m}] \subseteq K$ and therefore also that $K[r_K(\boldsymbol{m})] = A[\boldsymbol{m}]$. We also infer from Lemma 44 that $K \in \min S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$, which tells us that $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$. Since $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} = p_{|\Sigma_K}\mathrm{B}_{A,\boldsymbol{m}|\Sigma_K} = p_{|\Sigma_K}\mathrm{B}_{K,r_K(\boldsymbol{m})}$, we find that $p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$.

Next, we prove that $\mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}} \subseteq \mathcal{H}^s_{\mathrm{SI},A}\rfloor\boldsymbol{m}$. Consider any $p \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$, meaning that $p \neq 0$ and $p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ for all $K \in \min S_{A,\boldsymbol{m}}(p)$. We set out to prove that $p\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{SI},A}$ or, equivalently, that $p\mathrm{B}_{A,\boldsymbol{m}} \neq 0$ and that $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$ for all $K \in \min S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$. Since $p \neq 0$, the continuity of polynomials guarantees that there is some $\boldsymbol{\theta} \in \mathrm{int}(\Sigma_A)$ such that $p(\boldsymbol{\theta}) \neq 0$ and therefore also $(p\mathrm{B}_{A,\boldsymbol{m}})(\boldsymbol{\theta}) \neq 0$. So we know already that $p\mathrm{B}_{A,\boldsymbol{m}} \neq 0$. Consider any $K \in \min S_{A,\boldsymbol{0}}(p\mathrm{B}_{A,\boldsymbol{m}})$. Then, by Lemma 44, $K \in \min S_{A,\boldsymbol{m}}(p)$, implying that $A[\boldsymbol{m}] \subseteq K$ and $p_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ and therefore $p_{|\Sigma_K}\mathrm{B}_{K,r_K(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},K}$. Since moreover $p_{|\Sigma_K}\mathrm{B}_{K,r_K(\boldsymbol{m})} = (p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K}$, we find that indeed $(p\mathrm{B}_{A,\boldsymbol{m}})_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}$. $\square$

Proof of Equation (63). Due to Equation (17), it suffices to prove that, for any $p \in \mathcal{V}(A)$, $\min_{x\in A} p(\boldsymbol{\theta}^\circ_x) > 0 \Rightarrow p \in \mathcal{H}^s_{\mathrm{SI},A}$ and $\min_{x\in A} p(\boldsymbol{\theta}^\circ_x) < 0 \Rightarrow p \notin \mathcal{H}^s_{\mathrm{SI},A}$.

First, assume that $\min_{x\in A} p(\boldsymbol{\theta}^\circ_x) < 0$. Then there is some $y \in A$ for which $p(\boldsymbol{\theta}^\circ_y) < 0$. Hence, since $p_{|\Sigma_{\{y\}}} = p(\boldsymbol{\theta}^\circ_y) < 0$, we find that $\{y\} \in \min S_{A,\boldsymbol{0}}(p)$ and therefore also, due to the Bernstein coherence of $\mathcal{H}^s_{\mathrm{IDM},\{y\}}$ [see Theorem 21], that $p_{|\Sigma_{\{y\}}} \notin \mathcal{H}^s_{\mathrm{IDM},\{y\}}$, from which we infer that $p \notin \mathcal{H}^s_{\mathrm{SI},A}$.

Next, assume that $\min_{x\in A} p(\boldsymbol{\theta}^\circ_x) > 0$. Then $p_{|\Sigma_{\{x\}}} = p(\boldsymbol{\theta}^\circ_x) > 0$ for all $x \in A$, implying that $\min S_{A,\boldsymbol{0}}(p) = \{\{x\} \colon x \in A\}$ and that, for all $x \in A$, $p_{|\Sigma_{\{x\}}} \in \mathcal{H}^s_{\mathrm{IDM},\{x\}}$, again because of the Bernstein coherence of $\mathcal{H}^s_{\mathrm{IDM},\{x\}}$. Hence, since $p \neq 0$, we find that $p \in \mathcal{H}^s_{\mathrm{SI},A}$. $\square$

Proof of Equations (64) *and* (65). Equation (65) follows directly from Equation (55). We prove Equation (64). Due to Equation (17) and Proposition 23, it suffices to prove that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and any $p \in \mathcal{V}(A)$:

$$c(p, \check{\boldsymbol{m}}) > 0 \Rightarrow p \in \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}} \text{ and } c(p, \check{\boldsymbol{m}}) < 0 \Rightarrow p \notin \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}},$$

where, for ease of notation, we let

$$c(p, \check{\boldsymbol{m}}) := \inf_{\boldsymbol{\alpha}\in\Delta^s_{A[\check{\boldsymbol{m}}]}} \mathrm{Di}_{A[\check{\boldsymbol{m}}]}(p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} | r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}}) + \boldsymbol{\alpha}).$$

First, assume that $c(p,\check{\boldsymbol{m}}) < 0$, implying that $\mathrm{Di}_{A[\check{\boldsymbol{m}}]}(p_{|\Sigma_{A[\check{\boldsymbol{m}}]}}|r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}})+\boldsymbol{\alpha}) < 0$ for some $\boldsymbol{\alpha} \in \Delta^s_{A[\check{\boldsymbol{m}}]}$ and therefore also that $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \neq 0$ and, by Equation (54), that $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \notin \mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}\rfloor r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}})$. Hence, we find that $A[\check{\boldsymbol{m}}] \in \min S_{A,\check{\boldsymbol{m}}}(p)$ and therefore also that $p \notin \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}}$.

Next, assume that $c(p,\check{\boldsymbol{m}}) > 0$, implying that $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \neq 0$ and, by Equation (54), that $p_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \in \mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}\rfloor r_{A[\check{\boldsymbol{m}}]}(\check{\boldsymbol{m}})$. Hence, we find that $\min S_{A,\check{\boldsymbol{m}}}(p) = \{A[\check{\boldsymbol{m}}]\}$ and therefore also that $p \in \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}}$. □

Proof of Equation (67)*.* By combining Equations (62) and (30), we see that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\mathcal{D}^{s,1}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{L}(A)\colon\ \mathrm{S}_A(f) \in \mathcal{H}^s_{\mathrm{SI},A,\check{\boldsymbol{m}}}\right\}. \tag{96}$$

Consider now any $f \in \mathcal{L}(A)$ and distinguish between two cases: $f_{|A[\check{\boldsymbol{m}}]} \neq 0$ and $f_{|A[\check{\boldsymbol{m}}]} = 0$.

If $f_{|A[\check{\boldsymbol{m}}]} \neq 0$ [and therefore also $f \neq 0$], then

$$\begin{aligned}
f \in \mathcal{D}^{s,1}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}} &\Leftrightarrow \mathrm{S}_A(f)_{|\Sigma_{A[\check{\boldsymbol{m}}]}} \in \mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}\rfloor r_{A[\check{\boldsymbol{m}}]}(\boldsymbol{m})\\
&\Leftrightarrow \mathrm{S}_{A[\check{\boldsymbol{m}}]}(f_{|A[\check{\boldsymbol{m}}]}) \in \mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}\rfloor r_{A[\check{\boldsymbol{m}}]}(\boldsymbol{m})\\
&\Leftrightarrow f_{|A[\check{\boldsymbol{m}}]} \in \mathcal{D}^{s,1}_{\mathrm{IDM},A[\check{\boldsymbol{m}}]}\rfloor r_{A[\check{\boldsymbol{m}}]}(\boldsymbol{m})\\
&\Leftrightarrow f_{|A[\check{\boldsymbol{m}}]} > -\frac{1}{s}\sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\check{m}_x \Leftrightarrow f_{|A[\check{\boldsymbol{m}}]} > -\frac{1}{s}\sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\check{m}_x \text{ or } f > 0,
\end{aligned}$$

where the first equivalence is due to Statement (93) and Equations (94), (61) and (96). The second equivalence follows from the definition of S_A and $\mathrm{S}_{A[\check{\boldsymbol{m}}]}$ and the third one is due to Equations (21) and (30). The fourth equivalence is a consequence of Equation (67) and the final equivalence holds because $f > 0$ is redundant, given that $f_{|A[\check{\boldsymbol{m}}]} \neq 0$.

If $f_{|A[\check{\boldsymbol{m}}]} = 0$, then [again, using Statement (93) and Equations (94), (61) and (96)] $f \in \mathcal{D}^{s,1}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}}$ if and only if $f \neq 0$ and if for all $x \in A \setminus A[\check{\boldsymbol{m}}]$:

$$f(x) = 0 \text{ or } \mathrm{S}_A(f)_{|\Sigma_{A[\check{\boldsymbol{m}}]\cup\{x\}}} \in \mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]\cup\{x\}}\rfloor r_{A[\check{\boldsymbol{m}}]\cup\{x\}}(\boldsymbol{m}).$$

Since $f_{|A[\check{\boldsymbol{m}}]} = 0$ and $\mathcal{H}^s_{\mathrm{IDM},A[\check{\boldsymbol{m}}]\cup\{x\}}\rfloor r_{A[\check{\boldsymbol{m}}]\cup\{x\}}(\boldsymbol{m})$ is Bernstein coherent [Theorem 21], the latter statement is equivalent to $f(x) > 0$. Hence, we find that:

$$\begin{aligned}
f \in \mathcal{D}^{s,1}_{\mathrm{SI},A}\rfloor\check{\boldsymbol{m}} &\Leftrightarrow f \neq 0 \text{ and } (\forall x \in A \setminus A[\check{\boldsymbol{m}}]) f(x) \geq 0\\
&\Leftrightarrow f > 0\\
&\Leftrightarrow f_{|A[\check{\boldsymbol{m}}]} > -\frac{1}{s}\sum_{x\in A[\check{\boldsymbol{m}}]} f(x)\check{m}_x \text{ or } f > 0,
\end{aligned}$$

where the second and third equivalences are consequences of $f_{|A[\check{\boldsymbol{m}}]} = 0$. □

Lemma 55. *Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, and any $\emptyset \neq K \subseteq A$ such that $A[\boldsymbol{m}] \subseteq K$. Then $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m}) \Leftrightarrow p_{|\Sigma_{\rho(K)}} \in \mathcal{H}^s_{\mathrm{IDM},\rho(K)}\rfloor r_{\rho(K)}(R_\rho(\boldsymbol{m}))$.*

Proof of Lemma 55. Let $A^* := K$, $D^* := \rho(K)$, $\rho^* := \rho_{|K}$ and $p^* := p_{|\Sigma_{\rho(K)}}$. Then ρ^* is an onto map from A^* to D^*, $p^* \in \mathcal{V}(D^*)$ and $p^* \circ R_{\rho^*} = p_{|\Sigma_{\rho(K)}} \circ R_{\rho_{|K}} = (p \circ R_\rho)_{|\Sigma_K}$. Since $A[\boldsymbol{m}] \subseteq K$, we can identify $\boldsymbol{m}$ with an element $\boldsymbol{m}^* := r_K(\boldsymbol{m})$ of $\mathcal{N}_K^{m_A}$ and therefore the result follows from the representation insensitivity of the IDMM inference system with hyperparameter s, because then also $R_{\rho^*}(\boldsymbol{m}^*) = R_{\rho_{|K}}(r_K(\boldsymbol{m})) = r_{\rho(K)}(R_\rho(\boldsymbol{m}))$:

$$p^* \circ R_{\rho^*} \in \mathcal{H}^s_{\mathrm{IDM},A^*}\rfloor \boldsymbol{m}^* \Leftrightarrow p^* \in \mathcal{H}^s_{\mathrm{IDM},D^*}\rfloor R_{\rho^*}(\boldsymbol{m}^*). \qquad \square$$

Proposition 56. Φ^s_{SI} *is representation insensitive.*

Proof of Proposition 56. We use the characterisation of representation insensitivity in Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then, by Proposition 23, we need to prove that $p \circ R_\rho \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}} \Leftrightarrow p \in \mathcal{H}^s_{\mathrm{SI},D,R_\rho(\boldsymbol{m})}$.

First, assume that $p \in \mathcal{H}^s_{\mathrm{SI},D,R_\rho(\boldsymbol{m})}$, meaning that $p \neq 0$ and $p_{|\Sigma_L} \in \mathcal{H}^s_{\mathrm{IDM},L}\rfloor r_L(R_\rho(\boldsymbol{m}))$ for all $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$. Applying Lemma 45 with $K = A$, we infer from $p \neq 0$ that $p \circ R_\rho \neq 0$. Consider any $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$, so $\emptyset \neq K \subseteq A$ and $A[\boldsymbol{m}] \subseteq K$. Then, by Lemma 46, $\rho(K) \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$, implying that, due to the assumption, $p_{|\Sigma_{\rho(K)}} \in \mathcal{H}^s_{\mathrm{IDM},\rho(K)}\rfloor r_{\rho(K)}(R_\rho(\boldsymbol{m}))$. Applying Lemma 55, we find that $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$. Hence, $p \circ R_\rho \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$.

Assume, conversely, that $p\circ R_\rho \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$, meaning that $p\circ R_\rho \neq 0$ and that $(p\circ R_\rho)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ for all $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$. Applying Lemma 45 with $K = A$, we infer from $p \circ R_\rho \neq 0$ that $p \neq 0$. Consider any $L \in \min S_{D,R_\rho(\boldsymbol{m})}(p)$. By Lemma 46, there is some $K \in \min S_{A,\boldsymbol{m}}(p \circ R_\rho)$ such that $\rho(K) = L$. Since $\emptyset \neq K \subseteq A$, $A[\boldsymbol{m}] \subseteq K$ and, by the assumption, $(p \circ R_\rho)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$, we infer from Lemma 55 that $p_{|\Sigma_L} \in \mathcal{H}^s_{\mathrm{IDM},L}\rfloor r_L(R_\rho(\boldsymbol{m}))$. Hence, $p \in \mathcal{H}^s_{\mathrm{SI},D,R_\rho(\boldsymbol{m})}$. $\square$

Lemma 57. *Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, any $r \in \mathbb{N}_0$ and any $K \subseteq A$ such that $K \cap B \neq \emptyset$ and $A[\boldsymbol{m}] \subseteq K$. Then $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m}) \Leftrightarrow p_{|\Sigma_{K\cap B}} \in \mathcal{H}^s_{\mathrm{IDM},K\cap B}\rfloor r_{K\cap B}(\boldsymbol{m})$.*

Proof of Lemma 57. Let $A^* := K$, $B^* := K \cap B$, $p^* := p_{|\Sigma_{K\cap B}}$ and $r^* = \deg(p) - \deg(p^*) + r$. Then $B^* \subseteq A^*$, $p^* \in \mathcal{V}(B^*)$, $r^* \geq r \geq 0$, $r^* + \deg(p^*) = r + \deg(p)$, and

$$\begin{aligned}
\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} &= \sum_{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}} b_p^{\deg(p)+r}(\boldsymbol{n}) \mathrm{B}_{A,i_A(\boldsymbol{n})|\Sigma_K} = \sum_{\substack{\boldsymbol{n}\in\mathcal{N}_B^{\deg(p)+r}\\ B[\boldsymbol{n}]\subseteq K}} b_p^{\deg(p)+r}(\boldsymbol{n}) \mathrm{B}_{A,i_A(\boldsymbol{n})|\Sigma_K}\\
&= \sum_{\boldsymbol{n}\in\mathcal{N}_{K\cap B}^{\deg(p)+r}} b_{p_{|\Sigma_{K\cap B}}}^{\deg(p)+r}(\boldsymbol{n}) \mathrm{B}_{K,i_K(\boldsymbol{n})} = \mathrm{I}^{r^*}_{B^*,A^*}(p^*),
\end{aligned}$$

where the third equality follows from the unicity of the Bernstein expansion of a polynomial. Since $A[\boldsymbol{m}] \subseteq K$, we can identify $\boldsymbol{m}$ with an element $\boldsymbol{m}^* := r_K(\boldsymbol{m})$ of $\mathcal{N}_K \cup \{\boldsymbol{0}\}$ and therefore the result follows from the specificity of the IDMM inference system with hyperparameter s, because then also $r_{B^*}(\boldsymbol{m}^*) = r_{K\cap B}(\boldsymbol{m})$:

$$\mathrm{I}^{r^*}_{B^*,A^*}(p^*) \in \mathcal{H}^s_{\mathrm{IDM},A^*}\rfloor \boldsymbol{m}^* \Leftrightarrow p^* \in \mathcal{H}^s_{\mathrm{IDM},B^*}\rfloor r_{B^*}(\boldsymbol{m}^*). \qquad \square$$

Proposition 58. Φ^s_{SI} *is specific.*

Proof of Proposition 58. We use the characterisation of specificity in Theorem 11. Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $r \in \mathbb{N}_0$. Then, by Proposition 23, we need to prove that $\mathrm{I}^r_{B,A}(p) \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}} \Leftrightarrow p \in \mathcal{H}^s_{\mathrm{SI},B,r_B(\boldsymbol{m})}$. It is clear from Propositions 10 and 22 that we can assume without loss of generality that $r + \deg(p) > 0$.

First, assume that $p \in \mathcal{H}^s_{\mathrm{SI},B,r_B(\boldsymbol{m})}$, implying that $p \neq 0$ and that $p_{|\Sigma_L} \in \mathcal{H}^s_{\mathrm{IDM},L}\rfloor r_{B\cap L}(\boldsymbol{m})$ for all $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$. Applying Lemma 50 with $K = A$, we infer from $p \neq 0$ that $\mathrm{I}^r_{B,A}(p) \neq 0$. Consider any $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$. Then we infer from Lemma 51 that $K \cap B \in \min S_{B,r_B(\boldsymbol{m})}(p)$, implying that, due to our assumption, $p_{|\Sigma_{K\cap B}} \in \mathcal{H}^s_{\mathrm{IDM},K\cap B}\rfloor r_{K\cap B}(\boldsymbol{m})$. Since $K \cap B \neq 0$ and $A[\boldsymbol{m}] \subseteq K$, $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ because of Lemma 57. Hence, $\mathrm{I}^r_{B,A}(p) \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$.

Assume, conversely, that $\mathrm{I}^r_{B,A}(p) \in \mathcal{H}^s_{\mathrm{SI},A,\boldsymbol{m}}$, which implies that $\mathrm{I}^r_{B,A}(p) \neq 0$ and that $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$ for all $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$. Applying Lemma 50 with $K = A$, we infer from $\mathrm{I}^r_{B,A}(p) \neq 0$ that $p \neq 0$. Consider any $L \in \min S_{B,r_B(\boldsymbol{m})}(p)$. By Lemma 51, there is some $K \in \min S_{A,\boldsymbol{m}}(\mathrm{I}^r_{B,A}(p))$ such that $K \cap B = L$. Since $K \cap B \neq 0$, $A[\boldsymbol{m}] \subseteq K$ and, by assumption, $\mathrm{I}^r_{B,A}(p)_{|\Sigma_K} \in \mathcal{H}^s_{\mathrm{IDM},K}\rfloor r_K(\boldsymbol{m})$, we infer from Lemma 57 that $p_{|\Sigma_{K\cap B}} \in \mathcal{H}^s_{\mathrm{IDM},K\cap B}\rfloor r_{K\cap B}(\boldsymbol{m})$, or in other words, $p_{|\Sigma_L} \in \mathcal{H}^s_{\mathrm{IDM},L}\rfloor r_{B\cap L}(\boldsymbol{m})$. Hence, $p \in \mathcal{H}^s_{\mathrm{SI},B,r_B(\boldsymbol{m})}$. □

Proof of Theorem 24. This is an immediate consequence of Propositions 22 [coherence], 56 [representation insensitivity] and 58 [specificity]. □

E.9 Proofs of Results in Section 14

Proof of Theorem 25. We begin with coherence. Consider any category set A, then we have to prove that $\mathcal{H}_{\mathrm{H},A}$ is Bernstein coherent. For B1, recall that $0 \notin \mathcal{H}^s_{\mathrm{IDM},A}$ for all $s > 0$, and therefore also $0 \notin \mathcal{H}_{\mathrm{H},A}$. Similarly, for B2, recall that $\mathcal{V}^+(A) \subseteq \mathcal{H}^s_{\mathrm{IDM},A}$ for all $s > 0$, and therefore also $\mathcal{V}^+(A) \subseteq \mathcal{H}_{\mathrm{H},A}$. For B3, consider $n \in \mathbb{N}$ and $\lambda_k \in \mathbb{R}_{>0}$ and $p_k \in \mathcal{H}_{\mathrm{H},A}$ for all $k \in \{1,\ldots,n\}$. Then there is some $s > 0$ such that $p_k \in \mathcal{H}^s_{\mathrm{IDM},A}$ for all $k \in \{1,\ldots,n\}$, and therefore $\sum_{k=1}^n \lambda_k p_k \in \mathcal{H}^s_{\mathrm{IDM},A}$, by Bernstein coherence [Theorem 21]. Hence indeed $\sum_{k=1}^n \lambda_k p_k \in \mathcal{H}_{\mathrm{H},A}$.

Next, we turn to representation insensitivity, and use its characterisation in Theorem 7. Consider any category sets A and D such that there is an onto map $\rho\colon A \to D$, any $p \in \mathcal{V}(D)$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Then we find that indeed:

$$\begin{aligned}(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}_{\mathrm{H},A} &\Leftrightarrow (\exists s \in \mathbb{R}_{>0})(p \circ R_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{IDM},A}\\ &\Leftrightarrow (\exists s \in \mathbb{R}_{>0})p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},D} \Leftrightarrow p\mathrm{B}_{D,R_\rho(\boldsymbol{m})} \in \mathcal{H}_{\mathrm{H},D},\end{aligned}$$

where the second equivalence follows from the representation insensitivity of the IDMM inference systems [Theorem 21].

Finally, we turn to specificity, and use its characterisation in Theorem 11. Consider any category sets A and B such that $B \subseteq A$, any $p \in \mathcal{V}(B)$, any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $r \in \mathbb{N}_0$. Then we find that indeed:

$$\mathrm{I}^r_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}_{\mathrm{H},A} \Leftrightarrow (\exists s \in \mathbb{R}_{>0})\mathrm{I}^r_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \mathcal{H}^s_{\mathrm{IDM},A}$$

$$\Leftrightarrow (\exists s \in \mathbb{R}_{>0}) p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \mathcal{H}^s_{\mathrm{IDM},B} \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \mathcal{H}_{\mathrm{H},B},$$

where the second equivalence follows from the specificity of IDMM inference systems [Theorem 21]. □

Proof of Equation (69). For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $p \in \mathcal{V}(A)$:

$$\begin{aligned} p \in \mathcal{H}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} \Leftrightarrow p\mathrm{B}_{A,\check{\boldsymbol{m}}} \in \mathcal{H}_{\mathrm{H},A} &\Leftrightarrow (\exists s \in \mathbb{R}_{>0}) p\mathrm{B}_{A,\check{\boldsymbol{m}}} \in \mathcal{H}^s_{\mathrm{IDM},A} \\ &\Leftrightarrow (\exists s \in \mathbb{R}_{>0}) p \in \mathcal{H}^s_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}. \end{aligned}$$

Combined with Equation (54), this yields the desired result. □

Proof of Equation (71). For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and any $p \in \mathcal{V}(A)$:

$$\begin{aligned} \underline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) &= \sup\{\mu \in \mathbb{R} \colon p - \mu \in \mathcal{H}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}}\} \\ &= \sup_{s \in \mathbb{R}_{>0}} \sup\left\{\mu \in \mathbb{R} \colon p - \mu \in \mathcal{H}^s_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}}\right\} \\ &= \sup_{s \in \mathbb{R}_{>0}} \inf_{\boldsymbol{\alpha} \in \Delta^s_A} \mathrm{Di}_A(p|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) = \lim_{s \to +0} \inf_{\boldsymbol{\alpha} \in \Delta^s_A} \mathrm{Di}_A(p|\check{\boldsymbol{m}} + \boldsymbol{\alpha}), \end{aligned}$$

where the second equality is due to Equation (69), and the third one due to Equation (54). □

Proof of Equation (72). Consider any $p \in \mathcal{V}(A)$ and apply Equation (71):

$$\underline{H}_{\mathrm{H},A}(p) = \lim_{s \to +0} \inf_{\boldsymbol{\alpha} \in \Delta^s_A} \mathrm{Di}_A(p|\boldsymbol{\alpha}) = \lim_{s \to +0} \inf_{\boldsymbol{t} \in \mathrm{int}(\Sigma_A)} \inf_{s' \in (0,s)} \mathrm{Di}_A(p|s'\boldsymbol{t}) \tag{97}$$

Now fix any $n \geq \max\{\deg(p), 1\}$ and any $\boldsymbol{t} \in \mathrm{int}(\Sigma_A)$. Using Equation (80), we find that for all $\boldsymbol{m} \in \mathcal{N}^n_A$:

$$\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|s'\boldsymbol{t}) = \frac{1}{s'^{(n)}}\binom{n}{\boldsymbol{m}}\prod_{x \in A}(s't_x)^{(m_x)} = \frac{1}{s'^{(n)}}\binom{n}{\boldsymbol{m}}\prod_{x \in A[\boldsymbol{m}]}(s't_x)^{(m_x)},$$

where for all $x \in A[\boldsymbol{m}]$:

$$(s't_x)^{(m_x)} = (s't_x)(s't_x + 1)\dots(s't_x + m_x - 1) = s't_x(m_x - 1)![1 + O(s')]$$

and similarly:

$$\frac{1}{s'^{(n)}} = \frac{1}{s'(n-1)!}[1 + O(s')].$$

Hence, we find that

$$\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|s'\boldsymbol{t}) = \left(\frac{\prod_{x \in A[\boldsymbol{m}]} t_x(m_x - 1)!}{(n-1)!}\right) s'^{|A[\boldsymbol{m}]|-1}[1 + O(s')].$$

We now consider two cases: $|A[\boldsymbol{m}]| > 1$ and $|A[\boldsymbol{m}]| = 1$ [since $n \geq 1$, these cases are exhaustive]. If $|A[\boldsymbol{m}]| > 1$, then $\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|s'\boldsymbol{t}) = O(s')$. If $|A[\boldsymbol{m}]| = 1$ or, equivalently, if there is some $x \in A$ such that $\boldsymbol{m} = n\boldsymbol{e}^x$, then $\mathrm{Di}_A(\mathrm{B}_{A,n\boldsymbol{e}^x}|s'\boldsymbol{t}) = t_x[1+O(s')]$. If we combine this with Equation (78), we find that

$$\mathrm{Di}_A(p|s'\boldsymbol{t}) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\,\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{m}}|s'\boldsymbol{t}) = \sum_{x\in A} b_p^n(n\boldsymbol{e}^x)t_x + O(s').$$

Furthermore, again due to Equation (78):

$$b_p^n(n\boldsymbol{e}^x) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} b_p^n(\boldsymbol{m})\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}_x^\circ) = p(\boldsymbol{\theta}_x^\circ) \text{ for all } x \in A.$$

Hence, we conclude that

$$\mathrm{Di}_A(p|s'\boldsymbol{t}) = \sum_{x\in A} p(\boldsymbol{\theta}_x^\circ)t_x + O(s'),$$

which, combined with Equation (97), leads to the desired result. ☐

Proof of Equation (73). Consider any $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and any $p \in \mathcal{V}(A)$ and use Equation (71):

$$\underline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = \lim_{s\to+0} \inf_{\boldsymbol{\alpha}\in\Delta_A^s} \mathrm{Di}_A(p|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) \text{ and } \overline{H}_{\mathrm{H},A}(p|\check{\boldsymbol{m}}) = \lim_{s\to+0} \sup_{\boldsymbol{\alpha}\in\Delta_A^s} \mathrm{Di}_A(p|\check{\boldsymbol{m}}+\boldsymbol{\alpha}). \quad (98)$$

Since $\mathcal{H}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}}$ is Bernstein coherent [Theorem 25], it follows that $\underline{H}_{\mathrm{H},A}(\cdot|\check{\boldsymbol{m}})$ is a coherent lower prevision. This implies that $\underline{H}_{\mathrm{H},A}(\cdot|\check{\boldsymbol{m}})$ is super-additive, and that its conjugate upper prevision $\overline{H}_{\mathrm{H},A}(\cdot|\check{\boldsymbol{m}})$ is sub-additive. Hence, it suffices to prove the equalities in Equation (73) for any Bernstein basis polynomial $p = \mathrm{B}_{A,\boldsymbol{n}}$, where $\boldsymbol{n} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$. Now for any $\boldsymbol{\alpha} \in \Delta_A^s$ we gather from Equation (80) in Appendix B that:

$$\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{n}}|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) = \frac{1}{(\check{m}_A+\alpha_A)^{(n)}}\binom{n}{\boldsymbol{n}}\prod_{x\in A}(\check{m}_x+\alpha_x)^{(n_x)}.$$

Observe that:

$$(\check{m}_x+\alpha_x)^{(n_x)} = (\check{m}_x+\alpha_x)(\check{m}_x+\alpha_x+1)\dots(\check{m}_x+\alpha_x+n_x-1) = \check{m}_x^{(n_x)}[1+O(\alpha_x)],$$

and similarly, since $\check{m}_A > 0$:

$$\frac{1}{(\check{m}_A+\alpha_A)^{(n)}} = \frac{1}{\check{m}_A^{(n)}}[1+O(\alpha_A)]$$

Therefore:

$$\mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{n}}|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) = \binom{n}{\boldsymbol{n}}\frac{\prod_{x\in A}\check{m}_x^{(n_x)}}{\check{m}_A^{(n)}}[1+O(\alpha_A)]\prod_{x\in A}[1+O(\alpha_x)],$$

which, using Equation (98), leads to:[46]

$$\underline{H}_{\mathrm{H},A}(\mathrm{B}_{A,\boldsymbol{n}}|\check{\boldsymbol{m}}) = \overline{H}_{\mathrm{H},A}(\mathrm{B}_{A,\boldsymbol{n}}|\check{\boldsymbol{m}}) = \binom{n}{\boldsymbol{n}}\frac{\prod_{x\in A}\check{m}_x^{(n_x)}}{\check{m}_A^{(n)}} = \mathrm{Di}_A(\mathrm{B}_{A,\boldsymbol{n}}|\check{\boldsymbol{m}}).$$ ☐

46. See footnote 35.

E.10 Proofs of Results in Section 15

Proof of Theorem 26. We have already argued above that there is a smallest such inference system Φ, and we shall denote its lower probability function by φ. First, assume that $n \geq 2$. If we denote $\Delta\varphi(n,k) \coloneqq \varphi(n,k+1) - \varphi(n,k)$, then it follows from the assumptions that

$$\Delta\varphi(n,k+1) \leq \Delta\varphi(n,k) \text{ for } 0 \leq k \leq n-2. \tag{99}$$

We are first going to prove by induction that this implies that

$$\varphi(n,k) \geq \frac{k}{n}\varphi(n,n) \text{ for } 0 \leq k \leq n. \tag{100}$$

Observe that this inequality holds trivially for $k=0$ [Theorem 14.L1]. So assume that the inequality holds for $k=\ell$, where $\ell \in \{0,\dots,n-1\}$. Then we must show that it also holds for $k=\ell+1$. Assume, *ex absurdo*, that it does not, and therefore

$$\varphi(n,\ell+1) < \frac{\ell+1}{n}\varphi(n,n) \leq \varphi(n,\ell) + \frac{1}{n}\varphi(n,n), \tag{101}$$

where the second inequality follows from the induction hypothesis. Now we also have that

$$\begin{aligned}\varphi(n,n) &= \varphi(n,\ell+1) + \sum_{m=\ell+1}^{n-1}\Delta\varphi(n,m) \leq \varphi(n,\ell+1) + (n-\ell-1)\Delta\varphi(n,\ell)\\ &< \frac{\ell+1}{n}\varphi(n,n) + \frac{n-\ell-1}{n}\varphi(n,n) = \varphi(n,n),\end{aligned}$$

where the first inequality follows from Equation (99), and the second from the first and second inequalities in Equation (101). This is a contradiction, which completes our proof by induction of (100).

We infer from (100), Theorem 14.L9 and assumption (76) that

$$\varphi(n,k) \geq \frac{k}{n}\frac{n}{n+s} = \frac{k}{n+s} \text{ for } 0 \leq k \leq n. \tag{102}$$

Also observe that this inequality holds trivially for $n \in \{0,1\}$. We then get for the predictive lower prevision $\underline{P}^1_A(h|\boldsymbol{m})$ of any gamble h on A:

$$\begin{aligned}\underline{P}^1_A(h|\boldsymbol{m}) = \min h + \underline{P}^1_A(h-\min h|\boldsymbol{m}) &\geq \min h + \sum_{x\in A}[h(x)-\min h]\underline{P}^1_A(\mathbb{I}_{\{x\}}|\boldsymbol{m})\\ &= \min h + \sum_{x\in A}[h(x)-\min h]\varphi(n,m_x)\\ &\geq \min h + \sum_{x\in A}[h(x)-\min h]\frac{m_x}{n+s} = \underline{P}^{s,1}_{\mathrm{IDM},A}(h|\boldsymbol{m}),\end{aligned}$$

where the first equality and the first inequality follow from the coherence [P5, P2 and P3] of $\underline{P}^1_A(\cdot|\boldsymbol{m})$, the second equality from representation insensitivity [Equation (33)], and the second inequality from Equation (102). For the converse inequality, observe that the IDMM inference system Φ^s_{IDM} is coherent, representation insensitive, and specific by Theorem 21, clearly has concave surprise, satisfies assumption (76), and therefore dominates the smallest such inference system. $\square$

References

Augustin, T., Coolen, F. P. A., De Cooman, G., & Troffaes, M. C. M. (Eds.). (2014). *Introduction to Imprecise Probabilities*. John Wiley & Sons.

Bernard, J.-M. (1997). Bayesian analysis of tree-structured categorized data. *Revue Internationale de Systémique*, *11*, 11–29.

Bernard, J.-M. (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, *39*, 123–150.

Bernard, J.-M. (2007). In personal conversation..

Boole, G. (1847, reprinted in 1961). *The Laws of Thought*. Dover Publications, New York.

Boole, G. (2004, reprint of the work originally published by Watts & Co., London, in 1952). *Studies in Logic and Probability*. Dover Publications, Mineola, NY.

Carnap, R. (1952). *The continuum of inductive methods*. The University of Chicago Press.

Cifarelli, D. M., & Regazzini, E. (1996). De Finetti's contributions to probability and statistics. *Statistical Science*, *11*, 253–282.

Couso, I., & Moral, S. (2011). Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, *52*(7), 1034–1055.

Cozman, F. G. (2013). Independence for full conditional probabilities: Structure, factorization, non-uniqueness, and bayesian networks. *International Journal Of Approximate Reasoning*, *54*(9), 1261–1278.

De Cooman, G., & Miranda, E. (2007). Symmetry of models versus models of symmetry. In Harper, W. L., & Wheeler, G. R. (Eds.), *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*, pp. 67–149. King's College Publications.

De Cooman, G., & Miranda, E. (2008a). The F. Riesz Representation Theorem and finite additivity. In Dubois, D., Lubiano, M. A., Prade, H., Gil, M. A., Grzegorzewski, P., & Hryniewicz, O. (Eds.), *Soft Methods for Handling Variability and Imprecision (Proceedings of SMPS 2008)*, pp. 243–252. Springer.

De Cooman, G., & Miranda, E. (2008b). Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, *138*(8), 2409–2432.

De Cooman, G., & Miranda, E. (2012). Irrelevant and independent natural extension for sets of desirable gambles.. *Journal of Artificial Intelligence Research*, *45*, 601–640.

De Cooman, G., Miranda, E., & Quaeghebeur, E. (2009a). Representation insensitivity in immediate prediction under exchangeability. *International Journal of Approximate Reasoning*, *50*(2), 204–216.

De Cooman, G., & Quaeghebeur, E. (2012). Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, *53*(3), 363–395. Special issue in honour of Henry E. Kyburg, Jr.

De Cooman, G., Quaeghebeur, E., & Miranda, E. (2009b). Exchangeable lower previsions. *Bernoulli*, *15*(3), 721–735.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*, 1–68. English translation by Kyburg Jr. and Smokler (1964).

de Finetti, B. (1970). *Teoria delle Probabilità*. Einaudi, Turin.

de Finetti, B. (1974–1975). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester. English translation de Finetti's (1970) book, two volumes.

Dubins, L. E. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, *3*, 88–99.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall.

Goldstein, M. (1983). The prevision of a prevision. *Journal of the American Statistical Society*, *87*, 817–819.

Goldstein, M. (1985). Temporal coherence. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., & Smith, A. F. M. (Eds.), *Bayesian Statistics*, Vol. 2, pp. 231–248. North-Holland, Amsterdam. With discussion.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. The MIT Press.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, *33*, 222–225.

Hausdorff, F. (1923). Momentprobleme für ein endliches Intervall. *Mathematische Zeitschrift*, *13*, 220–248.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1998). *Theory of Probability*. Oxford Classics series. Oxford University Press. Reprint of the third edition (1961), with corrections.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York.

Johnson, W. E. (1924). *Logic, Part III. The Logical Foundations of Science*. Cambridge University Press. Reprinted by Dover Publications in 1964.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan, London.

Koopman, B. O. (1940). The Axioms and Algebra of Intuitive Probability. *The Annals of Mathematics, Second Series*, *41*(2), 269–292.

Kyburg Jr., H. E., & Smokler, H. E. (Eds.). (1964). *Studies in Subjective Probability*. Wiley, New York. Second edition (with new material) 1980.

Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction*. John Wiley & Sons.

Levi, I. (1980). *The Enterprise of Knowledge*. MIT Press, London.

Mangili, F., & Benavoli, A. (2013). New prior near-ignorance models on the simplex. In Cozman, F., Denœux, T., Destercke, S., & Seidenfeld, T. (Eds.), *ISIPTA '13 – Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pp. 213–222. SIPTA.

Miranda, E. (2009). Updating coherent lower previsions on finite spaces. *Fuzzy Sets and Systems*, *160*(9), 1286–1307.

Miranda, E., & De Cooman, G. (2014). *Introduction to Imprecise Probabilities*, chap. Lower previsions. John Wiley & Sons.

Miranda, E., & Zaffalon, M. (2011). Notes on desirability and conditional lower previsions. *Annals Of Mathematics And Artificial Intelligence*, *60*(3-4), 251–309.

Moral, S. (2005). Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, *45*, 197–214.

Moral, S., & Wilson, N. (1995). Revision rules for convex sets of probabilities. In Coletti, G., Dubois, D., & Scozzafava, R. (Eds.), *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, pp. 113–128. Plenum Press, New York.

Piatti, A., Zaffalon, M., Trojani, F., & Hutter, M. (2009). Limits of learning about a categorical latent variable under prior near-ignorance. *International Journal Of Approximate Reasoning*, *50*(4), 597–611.

Prautzsch, H., Boehm, W., & Paluszny, M. (2002). *Bézier and B-Spline Techniques*. Springer, Berlin.

Quaeghebeur, E. (2014). *Introduction to Imprecise Probabilities*, chap. Desirability. John Wiley & Sons.

Quaeghebeur, E., De Cooman, G., & Hermans, F. (2014). Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning*. Accepted for publication.

Rouanet, H., & Lecoutre, B. (1983). Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, *36*(2), 252–268.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (1995). A representation of partially ordered preferences. *The Annals of Statistics*, *23*, 2168–2217. Reprinted in the collection by Seidenfeld et al. (1999, pp. 69–129).

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (1999). *Rethinking the Foundations of Statistics*. Cambridge University Press, Cambridge.

Sheffer, I. M. (1939). Some properties of polynomial sets of type zero. *Duke Mathematical Journal*, *5*, 590–622.

Smith, C. A. B. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series A*, *23*, 1–37.

Troffaes, M. C. M., & De Cooman, G. (2014). *Lower Previsions*. Wiley.

Trump, W., & Prautzsch, H. (1996). Arbitrary degree elevation of Bézier representations. *Computer Aided Geometric Design*, *13*, 387–398.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, *58*, 3–57. With discussion.

Walley, P. (1997). A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, *24*(4), 463–483.

Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, *24*, 125–148.

Walley, P., & Bernard, J.-M. (1999). Imprecise probabilistic prediction for categorical data. Tech. rep. CAF-9901, Laboratoire Cognition et Activitées Finalisées, Université de Paris 8.

Williams, P. M. (1975a). Coherence, strict coherence and zero probabilities. In *Proceedings of the Fifth International Congress on Logic, Methodology and Philosophy of Science*, Vol. VI, pp. 29–33. Dordrecht. Proceedings of a 1974 conference held in Warsaw.

Williams, P. M. (1975b). Notes on conditional previsions. Tech. rep., School of Mathematical and Physical Science, University of Sussex, UK. See also the revised journal version by Williams (2007).

Williams, P. M. (1976). Indeterminate probabilities. In Przelecki, M., Szaniawski, K., & Wojcicki, R. (Eds.), *Formal Methods in the Methodology of Empirical Sciences*, pp. 229–246. Reidel, Dordrecht. Proceedings of a 1974 conference held in Warsaw.

Williams, P. M. (2007). Notes on conditional previsions. *International Journal of Approximate Reasoning*, *44*, 366–383.

Zabell, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, *10*, 1090–1099. Reprinted in the collection by Zabell (2005).

Zabell, S. L. (2005). *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, Cambridge, UK.

Zaffalon, M., & Miranda, E. (2013). Probability and time. *Artificial Intelligence*, *198*, 1–51.