Representation insensitivity in immediate prediction under exchangeability [1]

Gert de Cooman [a,*] Enrique Miranda [b] and Erik Quaeghebeur [a,2]

[a] *Ghent University, SYSTeMS Research Group, Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium*

[b] *Rey Juan Carlos University, Dept. of Statistics and O.R., C-Tulipán, s/n, 28933 Móstoles, Spain*

Abstract

We consider immediate predictive inference, where a subject, using a number of observations of a finite number of exchangeable random variables, is asked to coherently model his beliefs about the next observation, in terms of a predictive lower prevision. We study when such predictive lower previsions are representation insensitive, meaning that they are essentially independent of the choice of the (finite) set of possible values for the random variables. We establish that such representation insensitive predictive models have very interesting properties, and show that among such models, the ones produced by the Imprecise Dirichlet-Multinomial Model are quite special in a number of ways. In the Conclusion, we discuss the open question as to how unique the predictive lower previsions of the Imprecise Dirichlet-Multinomial Model are in being representation insensitive.

Key words: Predictive inference, immediate prediction, Rule of Succession, lower prevision, imprecise probabilities, coherence, exchangeability, Representation Invariance Principle, representation insensitivity, Imprecise Dirichlet-Multinomial Model, Johnson's sufficientness postulate.
MSC: 60G25 60G09 60A99 62M20 62G99

* Corresponding author.

[1] This paper has been partially supported by the research grant G.0139.01 of the Flemish Fund for Scientific Research (FWO) and by the projects MTM2004-01269, TSI2004-06801-C04-01.

[2] Research financed by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Article published in IJAR 50 (2009) 204–216; doi: 10.1016/j.ijar.2008.03.010

1 Introduction

Consider a subject who is making $N > 0$ successive observations of a certain phenomenon. We represent these observations by N random variables $X_1, \ldots, X_N$. By *random variable*, we mean a variable about whose value the subject may entertain certain beliefs. We assume that at each successive instant k, the actual value of the random variables X_k can be determined in principle. To fix ideas, our subject might be looking for frogs in the Amazon forest, and then X_k is the species of the k-th frog he comes across. Or, he might, as an archetypical example, be drawing balls without replacement from an urn, in which case X_k could designate the color of the k-th ball taken from the urn.

In the type of predictive inference we consider here, our subject in some way uses zero or more observations $X_1, \ldots, X_n$ made previously, i.e., those up to a certain instant $n \in \{0, 1, \ldots, N-1\}$, to predict, or make inferences about, the values of the future, or as yet unmade, observations $X_{n+1}, \ldots, X_N$. Here, we only consider the problem of *immediate prediction*: he is only trying to predict, or make inferences about, the value of the next observation X_{n+1}.

We are particularly interested in the problem of making such predictive inferences under prior ignorance: initially, *before making any observation, our subject knows very little or nothing about what produces these observations*. In the urn example, this is the situation where he does not know the composition of the urn, e.g., how many balls there are, or what their colors are. What we do assume, however, is that our subject makes an assessment of *exchangeability* to the effect that the order in which a sequence of observations has been made does not matter for his predictions.

In such a situation, a subject usually determines, beforehand, a non-empty finite set $\mathscr{X}$ of possible values, also called *categories* for the random variables X_k. It is then sometimes held, especially by advocates of a logical interpretation to probability, that our subject's beliefs should be represented by some given family of predictive probability mass functions. Such a predictive family is made up of real-valued maps $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$, which give, for each $n = 0, \ldots, N-1$ and each $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathscr{X}^n$, the so-called *predictive* probability mass function for the $(n+1)$-th observation, given the values $(X_1, \ldots, X_n) = (x_1, \ldots, x_n) = \boldsymbol{x}$ of the n previous observations. Any such family should in particular reflect the above-mentioned exchangeability assessment. Cases in point are the Laplace–Bayes Rule of Succession in the case of two categories [1], or Carnap's more general λ-calculus [2].

The inferences in Carnap's λ-calculus, to give but one example, can strongly depend on the number of elements in the set $\mathscr{X}$. This may well be considered undesirable. If for instance, we consider drawing balls from an urn, predictive inferences about whether the next ball will be '*red or green*' ideally should not depend on whether we assume beforehand that the possible categories are '*red*', '*green*',

'*blue*' and '*any other color*', or whether we take them to be '*red or green*', '*blue*', '*yellow*' and '*any other color*'. This desirable property was called *representation invariance* by Peter Walley [3], who showed that it is satisfied by the so-called Imprecise Dirichlet-Multinomial Model (or IDMM for short [4]). The IDMM can be seen as a special system of predictive *lower previsions* and it is a (predictive) cousin of the parametric Imprecise Dirichlet Model (or IDM [3]). Lower previsions are behavioral belief models that generalize the more classical Bayesian ones, such as probability mass functions, or previsions. We assume that the reader is familiar with the basic aspects of the theory of coherent lower previsions [5]. Relatively short introductions can be found in papers by Walley [6] and by ourselves [7,8].

Here, we intend to study general systems of such predictive lower previsions. In Section 2, we give a general definition of such predictive systems and study a number of properties they can satisfy, such as coherence and exchangeability. In Section 3, we study the property of representation insensitivity for predictive systems, which is a stronger version of Walley's representation invariance, tailored to making inferences under prior ignorance. We show in Section 4 that there are representation insensitive and exchangeable predictive systems, by giving two examples. These two can be used to generate so-called mixing predictive systems, which are studied in Section 5. Among these mixing predictive systems, the ones corresponding to an IDMM take a special place, as they are the only ones to satisfy all the above-mentioned properties and an extra one, called *specificity*, related to behavior under conditioning. In the Conclusions, we list a number of interesting, as yet unresolved, questions. We have gathered proofs in an Appendix.

2 Predictive families and systems

2.1 Families of predictive lower previsions

First assume that, before the subject starts making the observations X_k, he fixes a non-empty and finite set $\mathscr{X}$ of possible values for all the random variables X_k. We now want to represent his beliefs about the value of the $(n+1)$-th observation X_{n+1}, if he came to observe the sequence of values $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n$ for the first n random variables, or in other words, if he came to know that $X_k = x_k$ for $k = 1, \ldots, n$. The model we propose for this is a lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on the set $\mathscr{L}(\mathscr{X})$ of all gambles on $\mathscr{X}$. Let us first make clear what this means.

A *gamble* f on $\mathscr{X}$ is a real-valued map on $\mathscr{X}$. It represents an uncertain reward, expressed in terms of some predetermined linear utility scale. When interpreted as a gamble on the outcome X_{n+1}, it yields a (possibly negative) reward of $f(x)$ utiles if the value of the next variable X_{n+1} turns out to be the category x in $\mathscr{X}$. The set of all gambles on $\mathscr{X}$ is denoted by $\mathscr{L}(\mathscr{X})$. The *lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$

of any gamble f on $\mathscr{X}$ is the subject's supremum acceptable price for buying this gamble, or in other words, the highest r such that he accepts the uncertain reward $f(X_{n+1}) - p$ for all $p < r$, conditional on his having observed the values $\boldsymbol{x} = (x_1, \ldots, x_n)$ for the first n variables $(X_1, \ldots, X_n)$. His corresponding *predictive upper prevision*, or infimum selling price for f, is then given by the conjugacy relationship: $\overline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = -\underline{P}^{n+1}_{\mathscr{X}}(-f|\boldsymbol{x})$.

A specific class of gambles is related to *events*, i.e., subsets A of $\mathscr{X}$. This is the class of indicators I_A that map any element of A to one and all other elements of $\mathscr{X}$ to zero. A lower prevision that is defined on (indicators of) events only is called a *lower probability*, and we often write $\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{x})$ instead of $\underline{P}^{n+1}_{\mathscr{X}}(I_A|\boldsymbol{x})$. The reader may wonder at this point why we work with the seemingly more complicated language of gambles and lower previsions, rather than with that of events and lower probabilities. The main reason is that, as Walley has shown [5], the former is much more expressive: in contradistinction with a coherent prevision, a coherent lower prevision is not completely characterized by the values it assumes on events.

By the *predictive lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, which models beliefs about the value of the next random variable X_{n+1} given the observations $(X_1, \ldots, X_n) = \boldsymbol{x}$, we mean the real-valued functional, defined on the set of all gambles $\mathscr{L}(\mathscr{X})$, that assigns to any gamble f its predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$. We assume that the subject has such a predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^n$ and all $n \in \{0, \ldots, N-1\}$, where $N > 0$ is some fixed positive integer, representing the maximum or total number of observations we are interested in. For $n = 0$, there is some slight abuse of notation here, because we then actually have an unconditional predictive lower prevision $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$ for the first observation X_1, and no observations have yet been made. We are thus led to the following definition.

Definition 1 (Family of predictive lower previsions) *Consider a finite and non-empty set of categories $\mathscr{X}$. An $\mathscr{X}$-*family of predictive lower previsions*, or* predictive $\mathscr{X}$*-*family *for short, for up to $N > 0$ observations is a set of predictive lower previsions $\boldsymbol{\sigma}^N_{\mathscr{X}} := \left\{\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) \colon \boldsymbol{x} \in \mathscr{X}^n \text{ and } n = 0, \ldots, N-1\right\}$.*

It is useful to consider the special case, mentioned in the Introduction, and quite common in the literature, of a family of predictive lower previsions of which all members $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ are actually *linear or coherent previsions* $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, i.e, such that for each $n = 0, \ldots, N-1$ and $\boldsymbol{x} \in \mathscr{X}^n$ there is some predictive *(probability) mass function* $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$ such that $p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) \geq 0$ and $\sum_{z \in \mathscr{X}} p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) = 1$, and where for all gambles f on $\mathscr{X}$, $P^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = \sum_{z \in \mathscr{X}} f(z) p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x})$. Such linear previsions are the Bayesian belief models usually encountered in the literature (see for instance de Finetti's book [9]). We can use Bayes's rule to combine these predictive mass functions into unique *joint mass functions* $p^n_{\mathscr{X}}$ on $\mathscr{X}^n := \times_{i=1}^n \mathscr{X}$, given by

$$p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(x_1, \ldots, x_n) = \prod_{k=0}^{n-1} p^{k+1}_{\mathscr{X}}(x_{k+1}|x_1, \ldots, x_k),$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathscr{X}^n$ and $n = 1, \ldots, N$. This leads to unique corresponding linear previsions $P^n_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X}^n)$, the set of gambles g on $\mathscr{X}^n$, given by

$$P^n_{\mathscr{X}}(g) = \sum_{\boldsymbol{x} \in \mathscr{X}^n} g(\boldsymbol{x}) p^n_{\mathscr{X}}(\boldsymbol{x}). \tag{1}$$

For $n = N$, we call $P^N_{\mathscr{X}}$ the *joint linear prevision* associated with the given predictive family of linear previsions. It models beliefs about the values that the random variables $(X_1, \ldots, X_N)$ assume *jointly* in $\mathscr{X}^N$.

2.2 Systems of predictive lower previsions

When a subject is using a family of predictive lower previsions $\sigma^N_{\mathscr{X}}$, this means he has assumed beforehand that the random variables X_1, …, X_N all take values in the set $\mathscr{X}$. It cannot, therefore, be excluded at this point that his inferences, as represented by the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, strongly depend on the choice of the set of possible values $\mathscr{X}$. Any initial choice of $\mathscr{X}$ may lead to an essentially very different predictive family $\sigma^N_{\mathscr{X}}$. In order to be able to deal with this possible dependence mathematically, we now define predictive systems as follows.

Definition 2 (System of predictive lower previsions) *Fix $N > 0$. Consider for any finite non-empty set of categories $\mathscr{X}$ an $\mathscr{X}$-family $\sigma^N_{\mathscr{X}}$ of predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$. The set $\sigma^N := \left\{\sigma^N_{\mathscr{X}} : \mathscr{X} \text{ is a finite and non-empty set}\right\}$ is called a* system of predictive lower previsions*, or* predictive system *for short, for up to N observations. We denote by Σ^N the set of all such predictive systems.*

It is such predictive systems whose properties we intend to study. For two predictive systems σ^N and λ^N we say that σ^N is *less committal*, or *more conservative*, than λ^N, and we denote this by $\sigma^N \preceq \lambda^N$, if each predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in σ^N is *point-wise dominated* by the corresponding predictive lower prevision $\underline{Q}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in λ^N: $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) \leq \underline{Q}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ for all gambles f on $\mathscr{X}$. The reason for this terminology should be clear: a subject using predictive system λ^N will be buying gambles f on $\mathscr{X}$ at supremum prices $\underline{Q}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ that are at least as high as the supremum prices $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ of a subject using predictive system σ^N.

The binary relation $\preceq$ on Σ^N is a partial order: it is reflexive, anti-symmetric and transitive. A non-empty subset $\left\{\sigma^N_\gamma : \gamma \in \Gamma\right\}$ of Σ^N may have an infimum (or greatest lower bound) with respect to this partial order, and whenever it exists, this infimum corresponds to taking *lower envelopes*: if we fix $\mathscr{X}$, n and $\boldsymbol{x}$, then the corresponding predictive lower prevision in the infimum predictive system is the lower envelope $\inf_{\gamma\in\Gamma} \underline{P}^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ of the corresponding predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ in the predictive systems σ^N_γ, $\gamma \in \Gamma$.

2.3 Coherence requirements

As is usually done for belief models, we impose certain consistency, or rationality, requirements on the members $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ of a predictive system $\boldsymbol{\sigma}^N$.

Definition 3 (Coherence) *A system of predictive lower previsions is called* coherent *if it is the infimum of a collection of systems of predictive linear previsions.*

This is equivalent to requiring, for each choice of $\mathscr{X}$, that the conditional lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for $n = 0,\ldots,N-1$ and $\boldsymbol{x} \in \mathscr{X}^n$ should satisfy Walley's (joint) coherence condition. [3] Coherence is in the present context [4] also equivalent to requiring that the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ should be *(separately) coherent*, meaning that for each finite and non-empty set $\mathscr{X}$, $n = 0,\ldots,N-1$ and $\boldsymbol{x}$ in $\mathscr{X}^n$, $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ should satisfy, for all gambles f and g on $\mathscr{X}$ and all real $\lambda \geq 0$:

(C1) $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) \geq \inf f$ [accepting sure gains];
(C2) $\underline{P}^{n+1}_{\mathscr{X}}(f+g|\boldsymbol{x}) \geq \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) + \underline{P}^{n+1}_{\mathscr{X}}(g|\boldsymbol{x})$ [super-additivity];
(C3) $\underline{P}^{n+1}_{\mathscr{X}}(\lambda f|\boldsymbol{x}) = \lambda \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ [positive homogeneity].

2.4 Exchangeability and regular exchangeability

Next, we show how to formulate an assessment of *exchangeability* of the random variables $X_1, \ldots, X_N$ in terms of a system of predictive lower previsions. A subject would make such an assessment if he believed that the order in which these variables are observed is not important. Let us make this idea more precise.

We begin with the definition of exchangeability for a *precise* predictive system, i.e., a system of predictive linear previsions. For each choice of $\mathscr{X}$, the precise $\mathscr{X}$-family $\boldsymbol{\sigma}^N_{\mathscr{X}}$ has a unique joint linear prevision $P^N_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X}^N)$, defined by Eq. (1), which describes beliefs about what values the joint random variable $(X_1,\ldots,X_N)$ assumes in $\mathscr{X}^N$. *We then call the precise predictive system exchangeable if all the associated joint linear previsions $P^N_{\mathscr{X}}$ are.* Formally, consider the set of all permutations of $\{1,\ldots,N\}$. With any such permutation π we can associate a permutation of $\mathscr{X}^N$, also denoted by π, that maps any $\boldsymbol{x} = (x_1,\ldots,x_N)$ in $\mathscr{X}^N$ to $\pi\boldsymbol{x} := (x_{\pi(1)},\ldots,x_{\pi(N)})$. Similarly, with any gamble f on $\mathscr{X}^N$, we can consider

[3] See Walley's book [5]: Section 6.2 for separate coherence, Section 7.1.4 for (joint) coherence of conditional lower previsions, and Section K3 for Williams's Theorem. Since the random variables X_k are assumed to only take on a finite number of values, Walley's coherence condition here coincides with the one first suggested by Williams [10].

[4] This follows from our generalized Marginal Extension Theorem for random variables [11, Theorem 4]: for any random variables $X_1, \ldots, X_N$, any separately coherent conditional lower previsions $\underline{P}_1$, $\underline{P}_2(\cdot|X_1)$, $\ldots$, $\underline{P}_N(\cdot|X_1,\ldots,X_{N-1})$ are automatically (jointly) coherent.

the permuted gamble $\pi f := f \circ \pi$, or in other words $(\pi f)(\boldsymbol{x}) = f(\pi \boldsymbol{x})$. We then require that $P^N_{\mathscr{X}}(\pi f) = P^N_{\mathscr{X}}(f)$ for any such permutation π and any gamble f on $\mathscr{X}^N$. Equivalently, in terms of the joint mass function $p^N_{\mathscr{X}}$, we require that $p^N_{\mathscr{X}}(\pi \boldsymbol{x}) = p^N_{\mathscr{X}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^N$ and all permutations π. See de Finetti's work [9,12] for more details and discussion of exchangeability for linear previsions.

We adopt the following definition of exchangeability for general predictive systems.

Definition 4 (Exchangeability) *A system of predictive lower previsions is called* exchangeable *if it is the infimum of a collection of exchangeable systems of predictive linear previsions. We denote by $\langle \Sigma^N_{\mathrm{e}}, \preceq \rangle$ the set of all exchangeable predictive systems for up to N observations, with the same order $\preceq$ as defined on $\langle \Sigma^N, \preceq \rangle$.*

It follows at once from this definition that the infimum of any non-empty collection of exchangeable predictive systems is still exchangeable, as an infimum of infima (and therefore an infimum itself) of collections of exchangeable systems of predictive linear previsions. This means that the partially ordered set $\langle \Sigma^N_{\mathrm{e}}, \preceq \rangle$ is a complete semi-lattice [13, Sections 3.19–3.20]. We next turn to a stronger requirement, introduced mainly for reasons of mathematical convenience.

Definition 5 (Regular exchangeability) *A system of predictive lower previsions is called* regularly exchangeable *if it is the infimum of some collection σ^N_γ, $\gamma \in \Gamma$, of exchangeable systems of predictive linear previsions, where for all finite non-empty $\mathscr{X}$, all $\boldsymbol{x}$ in $\mathscr{X}^{N-1}$, and all γ in Γ, $p^{N-1}_{\mathscr{X},\gamma}(\boldsymbol{x}) = \prod_{k=0}^{N-2} p^{k+1}_{\mathscr{X},\gamma}(x_{k+1}|x_1,\ldots,x_k) > 0$.*

The term *regular* reminds of the notion of regular extension considered by Walley [5, Appendix J]. Regular exchangeability implies that every predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ is the lower envelope of the predictive linear previsions $P^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$, uniquely derived from the joint linear previsions $P^N_{\mathscr{X},\gamma}$ by applying Bayes's rule:

$$P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x}) = \frac{P^N_{\mathscr{X},\gamma}(f I_{\{\boldsymbol{x}\}\times\mathscr{X}^{N-n}})}{P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})}, \text{or equivalently } p^{n+1}_{\mathscr{X},\gamma}(z|\boldsymbol{x}) = \frac{p^{n+1}_{\mathscr{X},\gamma}(\boldsymbol{x},z)}{p^n_{\mathscr{X},\gamma}(\boldsymbol{x})},$$

since the probability $p^n_{\mathscr{X},\gamma}(\boldsymbol{x}) := P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})$ of the conditioning event is non-zero. All regularly exchangeable predictive systems are in particular also exchangeable and coherent. A *precise exchangeable predictive system* is regularly exchangeable if and only if $p^{N-1}_{\mathscr{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathscr{X}^{N-1}$ and all finite non-empty sets $\mathscr{X}$: regular exchangeability is a stricter requirement than exchangeability.

The number of times $T_z(\boldsymbol{x}) := |\{k \in \{1,\ldots,n\} : x_k = z\}|$ that a given category z in $\mathscr{X}$ has been observed in some sample $\boldsymbol{x} \in \mathscr{X}^n$ of length $0 \leq n \leq N$, is of special importance when there is regular exchangeability. Consider the *counting map* $\boldsymbol{T}_{\mathscr{X}}$ that maps samples $\boldsymbol{x}$ of length n to the $\mathscr{X}$-tuple $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ whose components are $T_z(\boldsymbol{x})$, $z \in \mathscr{X}$. $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ tells us how often each of the elements of $\mathscr{X}$ appears in the sample $\boldsymbol{x}$, and as $\boldsymbol{x}$ varies over $\mathscr{X}^n$, $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ assumes all values in the set of

count vectors $\mathscr{N}_{\mathscr{X}}^n := \{\boldsymbol{m} \in \mathbb{N}_0^{\mathscr{X}} : \sum_{z\in\mathscr{X}} m_z = n\}$. Here $\mathbb{N}_0$ denotes the set of non-negative integers (including zero). It is easy to see that any two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n have the same count vector $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$ if and only if there is some permutation π of $\{1,\ldots,n\}$ such that $\boldsymbol{y} = \pi\boldsymbol{x}$. This leads to the following result.

Proposition 1 *In a precise exchangeable predictive system σ^N, consider any finite non-empty set $\mathscr{X}$, $0 \leq n \leq N-1$, and $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathscr{X}^n$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$. Then $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y})$. And if $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y}) > 0$, then $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$.*

As an immediate corollary, we see that in any regularly exchangeable predictive system, the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ only depend on the sample $\boldsymbol{x}$ through its count vector $\boldsymbol{m} = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$: for any other sample $\boldsymbol{y}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y}) = \boldsymbol{m}$, it holds that $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$ and we use the notation $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ for $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in order to reflect this. In fact, from now on we only consider predictive systems—be they regularly exchangeable or not—for which the predictive lower previsions only depend on the observed samples through their count vectors, or in other words, for which the count vectors are *sufficient statistics*.

Regular exchangeability allows us to prove the following inequality, which has far-reaching consequences. We denote by $\boldsymbol{e}_z$ the count vector in $\mathscr{N}^1_{\mathscr{X}}$ whose z-component is one and all of whose other components are zero; it corresponds to the case where we have a single observation of a category z in $\mathscr{X}$.

Proposition 2 *In a regularly exchangeable predictive system σ^N, it holds for all finite and non-empty sets $\mathscr{X}$, all $0 \leq n \leq N-2$, all $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$ that $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) \geq \underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m})$. Here $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)$ denotes the gamble on $\mathscr{X}$ that assumes the value $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_z)$ in $z \in \mathscr{X}$.*

3 Representation invariance and representation insensitivity

We now turn to Walley's notion of representation invariance; see his IDM paper [3] for detailed discussion and motivation. Representation invariance could also, and perhaps preferably so, be called *pooling invariance*. Consider a set of categories $\mathscr{X}$, and a partition $\mathscr{S}$ of $\mathscr{X}$. Each element S of such a partition corresponds to a single new category, that consists of all the elements $x \in S$ being pooled, i.e., considered as one. Denote by $S(x)$ the unique element of the partition $\mathscr{S}$ that a category $x \in \mathscr{X}$ belongs to. So we consider S as a map from $\mathscr{X}$ to $\mathscr{S}$. If a gamble g on $\mathscr{X}$ does not differentiate between pooled categories, or in other words, is constant on the elements of $\mathscr{S}$, this means that there is some gamble f on $\mathscr{S}$ such that $g = f \circ S$. Similarly, with a sample $\boldsymbol{x} = (x_1,\ldots,x_n) \in \mathscr{X}^n$, there corresponds a sample $S\boldsymbol{x} := (S(x_1),\ldots,S(x_n)) \in \mathscr{S}^n$ of pooled categories. We can of course consider the partition $\mathscr{S}$ as a set of categories, and then representation invariance requires that $\underline{P}^{n+1}_{\mathscr{X}}(f \circ S|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{S}}(f|S\boldsymbol{x})$: for gambles that do not differentiate between

pooled categories, it should not matter whether we consider predictive inferences for the set of original categories $\mathscr{X}$, or for the set of pooled categories $\mathscr{S}$.

Besides pooling invariance, we can also require *renaming invariance*: as long as no confusion can arise, it should not matter for a subject's predictive inferences what names he gives to the different categories. This may seem too trivial to even mention, and as far as we know, it is always implicitly taken for granted in predictive inference. But it will be well to devote some attention to it here, in order to distinguish it from the category permutation invariance to be discussed shortly, with which it is easily confused if we do not pay proper attention. If we have a renaming bijection λ between a set of categories $\mathscr{X}$ and a set of renamed categories $\mathscr{Y}$, where we clearly distinguish between the elements of $\mathscr{X}$ and those of $\mathscr{Y}$, then with a gamble f on the set of renamed categories, there corresponds a gamble $f \circ \lambda$ on the set of original categories $\mathscr{X}$. Similarly, with a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of original categories, there corresponds a sample of renamed categories $\lambda \boldsymbol{x} := (\lambda(x_1), \ldots, \lambda(x_n))$. Clearly, we should then require that $\underline{P}_{\mathscr{X}}(f \circ \lambda | \boldsymbol{x}) = \underline{P}_{\mathscr{Y}}(f | \lambda \boldsymbol{x})$.

We have already stated in the Introduction that we are 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 $\mathscr{X}$ he has chosen. To formalize this idea, consider a permutation ϖ of the elements of $\mathscr{X}$.[5] With any gamble f on $\mathscr{X}$, there corresponds a permuted gamble $f \circ \varpi$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a permuted sample $\varpi \boldsymbol{x} := (\varpi(x_1), \ldots, \varpi(x_n))$. If a subject has no reason to distinguish between categories z and their images ϖz, this means that $\underline{P}^{n+1}_{\mathscr{X}}(f \circ \varpi | \boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(f | \varpi \boldsymbol{x})$. We call this property *category permutation invariance*.[6] Formally, it closely resembles renaming invariance, but whereas the latter is a trivial requirement, category permutation invariance can only be justified when our subject has no reason to distinguish between the categories, which may for instance happen when he is in a state of prior ignorance. To draw attention to the difference between the two in a somewhat loose manner: category permutation invariance allows confusion between new and old categories, something which renaming invariance carefully avoids.

We call *representation insensitivity* the combination of representation, 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 essentially insensitive to the choice of representation, i.e., category set. It is not difficult to see that representation insensitivity can be formally characterized as follows. Consider two non-empty and finite sets of categories $\mathscr{X}$

[5] This permutation ϖ of the elements of $\mathscr{X}$, or in other words of the *categories*, should be contrasted with the permutation π of the order of the observations, i.e., of the time set $\{1, \ldots, N\}$, considered in Section 2.4 in order to define exchangeability.

[6] This requirement is related to the notion of (weak) permutation invariance that two of us studied in much detail in a paper [7] dealing with symmetry in uncertainty modeling.

and $\mathscr{Y}$, and a so-called *relabeling map* $\rho\colon \mathscr{X} \to \mathscr{Y}$ that is *onto*, i.e., such that $\mathscr{Y} = \rho(\mathscr{X}) := \{\rho(x)\colon x \in \mathscr{X}\}$. Then with any gamble f on $\mathscr{Y}$ there corresponds a gamble $f \circ \rho$ on $\mathscr{X}$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a transformed sample $\rho\boldsymbol{x} := (\rho(x_1), \ldots, \rho(x_n))$ in $\mathscr{Y}^n$. *Representation insensitivity for immediate prediction then means that* $\underline{P}^{n+1}_{\mathscr{X}}(f \circ \rho|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{Y}}(f|\rho\boldsymbol{x})$.

3.1 Definition and basic properties

For any gamble f on a finite non-empty set of categories $\mathscr{X}$, its range $f(\mathscr{X}) := \{f(x)\colon x \in \mathscr{X}\}$ can again be considered as a set of categories, and f itself can be seen as a relabeling map. With any $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ there corresponds a count vector $\boldsymbol{m}^f$ in $\mathscr{N}^n_{f(\mathscr{X})}$ defined by $m^f_r := \sum_{f(x)=r} m_x$ for all r in $f(\mathscr{X})$. Clearly, if $\boldsymbol{x}$ is a sample with count vector $\boldsymbol{m}$, then the relabeled sample $f\boldsymbol{x} = (f(x_1), \ldots, f(x_n))$ has count vector $\boldsymbol{m}^f$. Representation insensitivity is then equivalent to the following requirement, which we take as its definition, because of its simplicity and elegance.

Definition 6 (Representation insensitivity) *A predictive system* σ^N *is* representation insensitive *if for all* $0 \leq n \leq N-1$*, all finite non-empty sets* $\mathscr{X}$ *and* $\mathscr{Y}$*, all* $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ *and* $\boldsymbol{m'} \in \mathscr{N}^n_{\mathscr{Y}}$*, and all gambles* f *on* $\mathscr{X}$ *and* g *on* $\mathscr{Y}$ *such that* $f(\mathscr{X}) = g(\mathscr{Y})$*, the following implication holds:* $\boldsymbol{m}^f = \boldsymbol{m'}^g \Rightarrow \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{\mathscr{Y}}(g|\boldsymbol{m'})$.

Clearly, a predictive system σ^N is representation insensitive if and only if for all finite and non-empty sets $\mathscr{X}$, all $0 \leq n \leq N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all $f \in \mathscr{L}(\mathscr{X})$:

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{f(\mathscr{X})}(\mathrm{id}_{f(\mathscr{X})}|\boldsymbol{m}^f), \tag{2}$$

where $\mathrm{id}_{f(\mathscr{X})}$ denotes the identity map (gamble) on $f(\mathscr{X})$. The predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ then depends on $f(\mathscr{X})$ and $\boldsymbol{m}^f$ only, and not directly on $\mathscr{X}$, f and $\boldsymbol{m}$. To put it more explicitly, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ *only depends on the values that* f *may assume, and on the number of times each value has been observed.*

We denote by $\Sigma^N_{\mathrm{e,ri}}$ the set of all exchangeable predictive systems that are representation insensitive. It is a subset of the class Σ^N_{e} of all exchangeable predictive systems, and it inherits the order $\preceq$. Clearly, taking (non-empty) infima preserves representation insensitivity, so $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$ is a complete semi-lattice as well. We shall see further on in Theorem 6 that these two structures have the same bottom (the vacuous representation insensitive and exchangeable predictive system).

We are interested in finding, and studying the properties of, predictive systems that are both exchangeable (and therefore coherent) *and* representation insensitive. We believe performing such a study to be quite important, and we report on our first attempts to do so in the rest of this paper.

With any predictive system σ^N, we can associate a map φ_{σ^N} defined on the subset $\{(n,m)\colon 0 \leq m \leq n \leq N-1\}$ of $\mathbb{N}_0^2$ by

$$\varphi_{\sigma^N}(n,m) := \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}\,|n-m,m).$$

Why this map is important, becomes clear if we look at predictive systems that are representation insensitive. Consider any proper event $\emptyset \neq A \subset \mathscr{X}$, then it follows by applying Eq. (2) with $f = I_A$, that

$$\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) = \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}\,|n-m_A,m_A),= \varphi_{\sigma^N}(n,m_A) \tag{3}$$

where $m_A := \sum_{z\in A} m_z$. So we see that in a representation insensitive predictive system, the lower probability $\varphi_{\sigma^N}(n,m)$ of observing an event (that is neither considered to be impossible nor necessary) does not depend on the embedding set $\mathscr{X}$ nor on the event itself, but only on the total number of previous observations n, and on the number of times m that the event has been observed before. Something similar holds for the upper probability of observing a non-trivial event: by conjugacy,

$$\overline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) = 1 - \underline{P}^{n+1}_{\mathscr{X}}(A^c|\boldsymbol{m}) = 1 - \varphi_{\sigma^N}(n,m_{A^c}) = 1 - \varphi_{\sigma^N}(n,n-m_A), \tag{4}$$

where A^c denotes the set-theoretic complement of the event A. This property [7] of representation insensitive predictive systems is reminiscent of *Johnson's sufficientness postulate* [16] (we use Zabell's terminology [17]), which requires that the probability that the next observation will be some category x is a function $f_x(n,m_x)$ that depends only on x, on the number of times m_x that this category x has been observed before, and on the total number of previous observations n. Representation insensitivity is stronger: it entails that the function φ_{σ^N} that 'corresponds to' the f_x is the same for all categories x in all possible finite and non-empty sets $\mathscr{X}$.

We call φ_{σ^N} the *lower probability function* of the predictive system σ^N. To alleviate the notational complexity, we suppress the index and simply write φ, whenever it is clear which predictive system we are talking about. Let us now consider any predictive system σ^N that is representation insensitive and exchangeable. We show in the next section that there are such predictive systems. But first we look at a number of interesting properties for the associated lower probability function φ.

Proposition 3 *Let $N > 0$ and let σ^N be a representation insensitive and coherent predictive system with lower probability function φ. Then*

[7] Another interesting property of the map φ_{σ^N} is that it completely determines the values of the predictive system on gambles for those predictive systems which have the additional property of 2-monotonicity. This is for instance the case of the mixing predictive systems we shall study in Section 5. A thorough and general study of the condition of 2-monotonicity for lower previsions can be found in [14,15].

1. φ *is* $[0,1]$*-bounded:* $0 \leq \varphi(n,k) \leq 1$ *for all* $0 \leq k \leq n \leq N-1$.
2. φ *is super-additive in its second argument:* $\varphi(n,k+\ell) \geq \varphi(n,k) + \varphi(n,\ell)$ *for all non-negative integers n, k and* ℓ *such that* $k+\ell \leq n \leq N-1$.
3. $\varphi(n,0) = 0$ *for all* $0 \leq n \leq N-1$.
4. $\varphi(n,k) \geq k\varphi(n,1)$ *for* $1 \leq k \leq n \leq N-1$, *and* $0 \leq n\varphi(n,1) \leq 1$ *for* $1 \leq n \leq N-1$.
5. φ *is non-decreasing in its second argument:* $\varphi(n,k+1) \geq \varphi(n,k)$ *for* $0 \leq k < n \leq N-1$.

If σ^N *is moreover regularly exchangeable, then*

6. $\varphi(n,k) \geq \varphi(n+1,k) + \varphi(n,k)[\varphi(n+1,k+1) - \varphi(n+1,k)]$ *for* $0 \leq k \leq n \leq N-2$.
7. φ *is non-increasing in its first argument:* $\varphi(n+1,k) \leq \varphi(n,k)$ *for* $0 \leq k \leq n \leq N-2$.
8. $\varphi(n,1) \geq \varphi(n+1,1)[1+\varphi(n,1)]$ *for* $1 \leq n \leq N-2$.
9. *Suppose that* $\varphi(n,1) > 0$ *and define* $s_n := \frac{1}{\varphi(n,1)} - n$ *for* $1 \leq n \leq N-1$. *Then* $s_n \geq 0$, $\varphi(n,1) = 1/(s_n+n)$ *and* s_n *is non-decreasing.*

The s_n that appear in this proposition will later, in Section 5.2, turn out to be constant (independent of the number of observations n) under special additional assumptions, and will there play the rôle of the hyper-parameter s in the IDMM.

In particular, these results, together with Eqs. (3) and (4), allow us to draw interesting and intuitively appealing conclusions about predictive lower and upper probabilities, which are valid in any representation insensitive and coherent predictive system: (i) the lower probability of observing an event that has not been observed before is zero, and the upper probability of observing an event that has always been observed before is one [Proposition 3.3]; and (ii) if the number of observations remains fixed, then both the lower and the upper probability of observing an event again do not decrease if the number of times the event has already been observed increases [Proposition 3.5]. In predictive systems that are moreover regularly exchangeable, we also see that (iii) if the number of times an event has been observed remains the same as the number of observations increases, then the lower probability for observing the event again does not increase [Proposition 3.7].

When the predictive system we consider consists solely of families of predictive linear previsions (apart perhaps from predictive lower previsions for dealing with zero previous observations), we can use the additivity of linear previsions, instead of the mere super-additivity of (separately) coherent lower previsions used previously, to get stronger versions of parts of Proposition 3.

Corollary 4 *Consider a representation insensitive and coherent predictive system* σ^N*, with a lower probability function* φ*, and such that all the predictive lower previsions* $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ *for* $0 < n \leq N-1$ *are linear previsions. Then the following*

statements hold for all $0 < n \leq N-1$*:*

1. $\varphi(n,k+\ell) = \varphi(n,k) + \varphi(n,\ell)$ *for all* $k,\ell \geq 0$ *such that* $k+\ell \leq n$*;*
2. $\varphi(n,k) = k\varphi(n,1)$ *for all* $0 \leq k \leq n$*.*

4 Are there representation insensitive and exchangeable predictive systems?

We have not yet proven that our notions of representation insensitivity and exchangeability for predictive system are compatible, or in other words, we do not know yet if there are any predictive systems that are both representation insensitive and exchangeable (let alone regularly so). We remedy this situation here by establishing the existence of two 'extreme' types of representation insensitive and exchangeable predictive systems, one of which is also regularly exchangeable.

Consider, for any predictive system σ^N that is both representation insensitive and exchangeable, the predictive lower previsions for $n = 0$. These are actually unconditional lower previsions $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$, modeling our beliefs about the first observation X_1, i.e., when no observations have yet been made. It follows right away from Proposition 3 and Eqs. (3) and (4) that for any proper subset A of $\mathscr{X}$, $\underline{P}^1_{\mathscr{X}}(A) = \varphi(0,0) = 0$. Since $\underline{P}^1_{\mathscr{X}}$ is assumed to be a (separately) coherent lower prevision, Proposition 5 below then guarantees that $\underline{P}^1_{\mathscr{X}}(f) = \min f$, for any gamble f on $\mathscr{X}$. So *all the* $\underline{P}^1_{\mathscr{X}}$ *in a representation insensitive and exchangeable predictive system must be so-called vacuous lower previsions.*[8] This means that there is no choice for the first predictions. It also means that *it is impossible to achieve representation insensitivity in any precise predictive system* (but see Theorem 7 further on for a predictive system that comes close).

Proposition 5 *Consider an arbitrary non-empty set* $\mathscr{X}$*. Let* $\underline{P}$ *be a coherent lower prevision on* $\mathscr{L}(\mathscr{X})$ *such that* $\underline{P}(A) = 0$ *for all* $A \subset \mathscr{X}$*. Then* $\underline{P}$ *is the* vacuous lower prevision *on* $\mathscr{X}$*, meaning that for all gambles* f *on* $\mathscr{X}$*,* $\underline{P}(f) = \inf f$*.*

This leads us to consider the so-called *vacuous* predictive system ν^N where all predictive lower previsions are vacuous: for all $0 \leq n \leq N-1$, all finite non-empty sets $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \min f$.

Theorem 6 *The vacuous predictive system* ν^N *is regularly exchangeable and representation insensitive. It is the bottom of the complete semi-lattice* $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$*. Its lower probability function is given by* $\varphi(n,m) = 0$ *for* $0 \leq m \leq n \leq N-1$*.*

[8] This result was proven, in another way, by Walley [5, Section 5.5.1], when he argued that his Embedding and Symmetry Principles under coherence only leave room for the vacuous lower prevision. In the special case that there are no prior observations ($n = 0$), the Embedding Principle is related to representation invariance, and the Symmetry Principle to what we have called category permutation invariance.

In the vacuous predictive system the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ do not depend on the number of observations n, nor on the observed count vectors $\boldsymbol{m}$. A subject who is using the vacuous predictive system is not learning anything from the observations. In particular, we see that representation insensitivity and (regular) exchangeability do not guarantee that we become more committal as we have more information at our disposal. Indeed, with the vacuous predictive system, whatever our subject has observed before, he always remains fully uncommittal. If we want a predictive system where something is really being learned from the data, it seems we need to make some 'leap of faith', and add something to our assessments that is not a mere consequence of exchangeability and representation insensitivity.

Are there less trivial examples of exchangeable and representation insensitive predictive systems? We know that we must make the vacuous choice for $n = 0$, but is there, for instance, a way to make the predictive lower previsions *precise*, or linear, for $n > 0$? The following theorem tells us there is only one such predictive system.

Theorem 7 *Consider a predictive system where for any $0 < n \leq N-1$ all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are actually linear previsions $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$. If this predictive system is representation insensitive, then*

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \sum_{z\in\mathscr{X}} f(z)\frac{m_z}{n} \tag{5}$$

for all $0 < n \leq N-1$, all finite non-empty sets $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. For its lower probability function φ, we then have $\varphi(n,k) = \frac{k}{n}$ for all $0 \leq k \leq n$ and $n > 0$. Moreover, the predictive previsions given by Eq. (5)*, together with the vacuous lower previsions for $n = 0$, constitute a representation insensitive and exchangeable (but not regularly so) predictive system π^N.*

We call the predictive system π^N described in Theorem 7 the *Haldane* predictive system. The name refers to the fact that a Bayesian inference model with a multinomial likelihood function using Haldane's (improper) prior (see, e.g., Jeffreys [18, p. 123]) would lead to these predictive previsions for $n > 0$. The fact that the lower probability function of Haldane predictive system is always $\varphi(n,k) = \frac{k}{n}$ for all $0 \leq k \leq n \leq N-1$ and $n > 0$, together with Corollary 4, implies that statements 6 and 8 in Proposition 3 hold with equality in this case. Moreover, we have $s_n = 0$ for all $n \geq 0$. Note that in this case the lower probability function coincides with the classical frequentist estimation: the (lower and upper) probability for an event that has been observed k times in n trials is equal to $\frac{k}{n}$.

It is an interesting consequence of Walley's Marginal Extension Theorem [5, Section 6.7.3] that for any finite and non-empty $\mathscr{X}$, the only joint lower prevision on $\mathscr{L}(\mathscr{X}^N)$ that is coherent with the Haldane predictive $\mathscr{X}$-family is given by $\underline{P}^N_{\mathscr{X}}(g) = \min_{z\in\mathscr{X}} g(z,\ldots,z)$ for all gambles g on $\mathscr{X}^N$.[9] The Haldane predictive

[9] This implies that the Haldane predictive system is not regularly exchangeable: any dom-

system only seems to be coherent with a joint lower prevision $\underline{P}^N_{\mathscr{X}}$ which expresses that our subject is certain that all variables X_k will assume *the same value*, but where he is completely ignorant about what that common value is.

This is related to another observation: we deduce from Proposition 3.3 that in the Haldane predictive system, when $n > 0$ then not only the lower probability but also the upper probability of observing an event that has not been observed before is zero! This models that a subject is practically certain (because prepared to bet at all odds on the fact) that any event that has not been observed in the past will not be observed in the future either. The *sampling prevision* $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ for a gamble f in this predictive system is the expectation of f with respect to the observed (sampling) probability distribution on the set of categories. The Haldane predictive system is too strongly tied to the observations, and does not allow us to make 'reasonable' inferences in a general context. The Haldane and the vacuous predictive systems are both extreme cases: in the latter the predictive lower previsions are independent of the observed data, and in the former they depend too strongly on them.

5 Mixing predictive systems

We have found two representation insensitive and exchangeable predictive systems, and both are not very useful: the first, because it does not allow learning from past observations, and the second, because its inferences are too strong and we seem to infer too much from the data. A natural question then is: can we find 'intermediate' representation insensitive and exchangeable predictive systems whose behavior is stronger than the vacuous and weaker than the Haldane predictive system? A simple way to get further models is to look at convex mixtures. Let us, therefore, consider a finite sequence $\boldsymbol{\varepsilon}$ of N numbers $\varepsilon_n \in [0,1]$, $0 \leq n \leq N-1$, and study the *mixing predictive system* $\boldsymbol{\sigma}^N_{\boldsymbol{\varepsilon}}$ whose predictive lower previsions are given by

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \varepsilon_n S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + (1-\varepsilon_n)\min f, \tag{6}$$

for all $0 \leq n \leq N-1$, all finite non-empty sets $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. As $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ is only defined for $n > 0$, and since representation insensitivity and coherence require that $\underline{P}^1_{\mathscr{X}}$ should be vacuous, we always let $\varepsilon_0 = 0$ implicitly. We call any such sequence $\boldsymbol{\varepsilon}$ a *mixing sequence*, and we denote by $\varphi_{\boldsymbol{\varepsilon}}$ the lower probability function of the corresponding mixing predictive system $\boldsymbol{\sigma}^N_{\boldsymbol{\varepsilon}}$.

We are mainly interested in finding mixing predictive systems that are representation insensitive and (regularly) exchangeable. The following proposition tells us

---

inating precise exchangeable predictive system satisfies $p^{N-1}_{\mathscr{X}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathscr{X}^{N-1}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{m} \neq (N-1)\boldsymbol{e}_z$ for all $z \in \mathscr{X}$, and for any such $\boldsymbol{x}$, the requirements for regular exchangeability cannot be satisfied.

that the only real issue lies with exchangeability. Its immediate proof is based on the simple observation that representation insensitivity is preserved by taking convex mixtures of any predictive systems.

Proposition 8 *For any mixing sequence ε, the predictive system σ_ε^N is still representation insensitive. Moreover, let $0 \leq k \leq n \leq N-1$. Then $\varphi_\varepsilon(n,k) = \varepsilon_n \frac{k}{n}$, and if $\varepsilon_n > 0$ then $s_n = n\frac{1-\varepsilon_n}{\varepsilon_n}$ and $\varepsilon_n = \frac{n}{n+s_n}$. In particular $\varphi_\varepsilon(n,1) = \varepsilon_n/n$ is the lower probability of observing a non-trivial event that has been observed once before in n trials, $\varepsilon_n = n\varphi_\varepsilon(n,1)$ is the lower probability $\varphi_\varepsilon(n,n)$ of observing a non-trivial event that has always been observed before (n out of n times), and $s_n = \frac{1-\varphi_\varepsilon(n,n)}{\varphi_\varepsilon(n,1)}$ is the ratio of the upper probability of observing an event that has never been observed before to the lower probability of observing a non-trivial event that has been observed once before, in n trials.*

We have already argued that, to get away from making vacuous inferences, and to be able to learn from observations, we need to make some 'leap of faith' and go beyond merely requiring exchangeability and representation insensitivity. One of the simplest ways to do so, it seems, is to specify the numbers $\varphi(n,1)$ for $n = 1,\ldots,N-1$, i.e., to *specify, beforehand, the lower probability of observing any non-trivial event that has been observed only once in n trials*. We can then ask for the most conservative representation insensitive predictive system that exhibits these lower probabilities. Interestingly, mixing predictive systems play this part:

Theorem 9 *Consider $N > 0$ and a mixing sequence ε. Let σ^N be a representation insensitive coherent predictive system such that its associated lower probability function φ satisfies $\varphi(n,1) \geq \varphi_\varepsilon(n,1) = \varepsilon_n/n$ for all $0 < n \leq N-1$. Then $\sigma_\varepsilon^N \preceq \sigma^N$.*

Mixing predictive systems have a special part in this theory, because they are quite simple, and in some sense most conservative. They are quite simple because, as Proposition 8 shows, all that is needed to specify them is the values $\varphi(n,1)$ of the lower probability function, or in other words, the lower probabilities that an event will occur that has been observed once in n observations. Theorem 9 shows they are the most conservative coherent and representation insensitive predictive systems with the given values for $\varphi(n,1)$. We shall see that there are mixing predictive systems with a non-trivial mixing sequence ε that are also regularly exchangeable. First, we establish a necessary condition on ε for this to be the case.

5.1 The regular exchangeability of mixing predictive systems

Consider any mixing sequence ε and the corresponding mixing predictive system σ_ε^N. Let us first derive a necessary condition that the ε_n should satisfy for the mixing predictive system to be regularly exchangeable. For the corresponding lower probability function φ_ε it holds by Proposition 8 that $\varphi_\varepsilon(n,k) = \varepsilon_n \frac{k}{n}$; if we substi-

tute this in the inequality of Proposition 3.8 we see that it is necessary for regular exchangeability that the ε_n should satisfy

$$\frac{\varepsilon_n}{n} \geq \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1,\ldots,N-1. \tag{7}$$

We deduce from this that if one ε_n is zero, then all of the subsequent ε_{n+k} are zero as well: if inferences are vacuous after $n>0$ observations, they should also remain vacuous after subsequent ones. Or, to put it more boldly, in regularly exchangeable mixing predictive systems, if we are going to learn at all from observations, we have to start doing so from the first observation.

5.2 Predictive inferences for the IDMM

To recover the immediate predictions of the IDMM, it is of particular interest to investigate for which types of mixing predictive systems, or in other words, for which mixing sequences ε, we generally have an equality rather than only an inequality in the condition of Proposition 2, i.e., for which

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m}), \tag{8}$$

for all finite and non-empty $\mathscr{X}$, all $0\leq n\leq N-1$, all $\boldsymbol{m}\in\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$, where the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are given by Eq. (6). Using the definition of $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ and the (separate) coherence [use (C6) in the Appendix] of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$, we find that this is equivalent to the condition

$$\frac{\varepsilon_n}{n} = \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1\ldots,N-1, \tag{9}$$

i.e., where we have the equality in (7). Clearly, one ε_n is zero if and only if all of them are, which leads to the vacuous predictive system ν^N. From Theorem 6, we know this vacuous system to be regularly exchangeable (and representation insensitive). If we assume on the other hand that $\varepsilon_n>0$ for $n=1,\ldots,N$, and let $\zeta_n := n/\varepsilon_n = n+s_n \geq 1$, then the above equality can be rewritten as $\zeta_{n+1}=\zeta_n+1$, which implies that there is some $s\geq 0$ such that $\zeta_n = n+s$, or equivalently, $s_n = s$ and consequently, for $n=0,1,\ldots,N-1$:

$$\varepsilon_n = \frac{n}{n+s}, \quad \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \frac{n}{n+s}S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + \frac{s}{n+s}\min f. \tag{10}$$

The predictive lower previsions in Eq. (10) are precisely the ones that can be associated with the Imprecise Dirichlet-Multinomial Model (or IDMM) with hyperparameter s [4, Section 4.1]. We call mixing predictive systems of this type IDMM-*predictive systems*. The vacuous predictive system corresponds to letting $s\to\infty$.

Theorem 10 *The vacuous predictive system, and the* IDMM*-predictive systems for $s > 0$ are regularly exchangeable and representation insensitive, and they are the only mixing predictive systems for which the equality* (8) *holds.*

Among the mixing predictive systems, the ones corresponding to the IDMM are also special in another way, which points to a very peculiar, but in our view intuitively appealing, property of predictive inferences produced by the IDMM. Indeed, assume that in addition to observing a count vector $\boldsymbol{m}$ of n observations, we come know in some way that the $(n+1)$-th observation will belong to a proper subset A of $\mathscr{X}$, and nothing else—we might suppose for instance that an observation of X_{n+1} has been made, but that it is imperfect, and only allows us to conclude that $X_{n+1} \in A$. Then we can ask what the updated beliefs are, i.e., what $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is. Since $\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) = \varepsilon_n m_A/n > 0$ if and only if $m_A > 0$ and $\varepsilon_n > 0$, let us assume that indeed $m_A > 0$ and $\varepsilon_n > 0$, in which case the requirements of coherence allows us to determine $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ uniquely, using the so-called Generalized Bayes Rule [5, Section 6.4] on the conditional lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$: $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is then the unique real μ such that

$$\underline{P}^{n+1}_{\mathscr{X}}(I_A[f-\mu]|\boldsymbol{m}) = 0. \tag{11}$$

We then have the following characterization of IDMM-predictive systems, where we denote by f_A the restriction of the gamble f to the set A, by $\boldsymbol{m}_A$ the A-tuple obtained from $\boldsymbol{m}$ by dropping the components that correspond to elements outside A. The sum of the components of $\boldsymbol{m}_A$ is m_A.

Theorem 11 (Specificity) *The* IDMM*-predictive systems with $s > 0$ are the only mixing predictive systems with all $\varepsilon_n > 0$, $n = 1,\ldots,N-1$ that satisfy the additional requirement*

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A) = \underline{P}^{m_A+1}_{A}(f_A|\boldsymbol{m}_A) \tag{12}$$

for all $n = 1,\ldots,N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$, all gambles f on $\mathscr{X}$ and all proper subsets A of $\mathscr{X}$ such that $m_A > 0$.

We find the so-called *specificity* property of inferences—the term was coined by Bernard [19], who first studied this property in the context of predictive inference—characterized by Eq. (12) to be quite peculiar. Indeed, suppose that you have observed n successive outcomes, leading to a count vector $\boldsymbol{m}$. If you know in addition that X_{n+1} belongs to A, then Eq. (12) tells you that *the updated value $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is the same as the one you would get by discarding all the* previous *observations producing values outside A, and in effect only retaining the m_A observations that were inside A!* It is as if knowing that the $(n+1)$-th observation belongs to A allows you to ignore all the previous observations that happened to lie outside A. This is intuitively appealing, because it means that if you know that the outcome of the next observation belongs to A, only the related behavior (the values of f on A and the previous observations of this set) matter for your prediction.

6 Conclusions

We have considered the problem of representation insensitivity in immediate prediction. We have defined predictive systems, and the properties we imposed (exchangeability and representation insensitivity) have led us to consider mixing predictive systems and more specifically, IDMM-predictive systems (also satisfying Eq. (12)). Much more work is needed, however, be able to draw a complete picture of the issue of representation insensitivity in predictive systems. Indeed, while doing research for this paper, we have come across a multitude of questions that we have not yet been able to answer, and we list only a few of them here: (i) Are there (regularly) exchangeable and representation insensitive predictive systems that are not mixing predictive systems? (ii) Related questions are: are there (regularly) exchangeable and representation insensitive predictive systems that, unlike the mixing systems, are not completely determined by the probabilities $\varphi(n,1)$ of observing an event that has been observed only once before in n observations; are there such predictive systems whose behavior on gambles, unlike that of mixing systems, is not completely determined by the lower probability function φ; and are there such predictive systems whose lower probability function φ, unlike that of mixing systems, is not additive in the sense that $\varphi(n,k+\ell) = \varphi(n,k) + \varphi(n,\ell)$? (iii) Are there (regularly) exchangeable and representation insensitive mixing predictive systems that are not of the IDMM-type. i.e., for which the equalities (8) and (9) are not satisfied? (iv) Are there (regularly) exchangeable, representation insensitive non-mixing predictive systems that satisfy Eq. (12)? (v) Can we arrive at stronger conclusions if we consider that the observations X_n make up an infinite exchangeable sequence? (vi) Can more definite answers be given if we consider the general, rather than the immediate, prediction problem?

Acknowledgments

We wish to thank Jean-Marc Bernard, Frank Coolen and Thomas Augustin for useful discussions and comments.

Appendix: Proofs of main results

We start by mentioning a few properties of (separately) coherent lower previsions $\underline{P}$ on $\mathscr{L}(\mathscr{X})$, i.e., lower previsions that satisfy (C1)–(C3). It is easy to check that (C1)–(C3) also imply, for all gambles f and g on $\mathscr{X}$, and all real μ:

(C4) $\underline{P}(f) \leq \sup f$;
(C5) $\underline{P}(f) \leq \underline{P}(g)$ if $f \leq g$ [monotonicity];

(C6) $\underline{P}(f+\mu)=\underline{P}(f)+\mu$ [constant additivity].

Proof of Proposition 1 We consider $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathscr{X}^n$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})=\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$. Then there is some permutation π of $\{1,\ldots,n\}$ such that $\boldsymbol{y}=\pi\boldsymbol{x}$, so it follows from exchangeability that $p^n_{\mathscr{X}}(\boldsymbol{y})=P^N_{\mathscr{X}}(\{\boldsymbol{y}\}\times\mathscr{X}^{N-n})=P^N_{\mathscr{X}}(\{\pi\boldsymbol{x}\}\times\mathscr{X}^{N-n})=P^N_{\mathscr{X}}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})=p^n_{\mathscr{X}}(\boldsymbol{x})$. We next assume $p^n_{\mathscr{X}}(\boldsymbol{x})=p^n_{\mathscr{X}}(\boldsymbol{y})>0$ to prove that $p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x})=p^{n+1}_{\mathscr{X}}(z|\boldsymbol{y})$ for all $z\in\mathscr{X}$. This follows immediately from the equalities $p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x})p^n_{\mathscr{X}}(\boldsymbol{x})=p^{n+1}_{\mathscr{X}}(\boldsymbol{x},z)=p^{n+1}_{\mathscr{X}}(\boldsymbol{y},z)=p^{n+1}_{\mathscr{X}}(z|\boldsymbol{y})p^n_{\mathscr{X}}(\boldsymbol{y})$, where the second equality again follows by applying exchangeability. □

Proof of Proposition 2 Consider any $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$, and any $\boldsymbol{x}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})=\boldsymbol{m}$. Regular exchangeability tells us that σ^N is the infimum of a collection $\sigma^N_\gamma,\gamma\in\Gamma$ of exchangeable precise predictive systems. Fix any γ in Γ and consider the corresponding exchangeable joint linear prevision $P^N_{\mathscr{X},\gamma}$. For any gamble f on $\mathscr{X}$, define the corresponding gambles g and g' on $\mathscr{X}^N$ by $g(\boldsymbol{z})=f(z_{n+1})I_{\{\boldsymbol{x}\}}(z_1,\ldots,z_n)$ and $g'(\boldsymbol{z})=f(z_{n+2})I_{\{\boldsymbol{x}\}}(z_1,\ldots,z_n)$ for all $\boldsymbol{z}=(z_1,\ldots,z_N)$ in $\mathscr{X}^N$. Observe that $P^N_{\mathscr{X},\gamma}(g)=P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x})p^n_{\mathscr{X},\gamma}(\boldsymbol{x})$ and that

$$\begin{aligned}P^N_{\mathscr{X},\gamma}(g')&=\sum_{(y_{n+1},y_{n+2})\in\mathscr{X}^2}f(y_{n+2})p^{n+2}_{\mathscr{X},\gamma}(y_{n+2}|\boldsymbol{x},y_{n+1})p^{n+1}_{\mathscr{X},\gamma}(y_{n+1}|\boldsymbol{x})p^n_{\mathscr{X},\gamma}(\boldsymbol{x})\\&=p^n_{\mathscr{X},\gamma}(\boldsymbol{x})\sum_{y_{n+1}\in\mathscr{X}}P^{n+2}_{\mathscr{X},\gamma}(f|\boldsymbol{x},y_{n+1})p^{n+1}_{\mathscr{X},\gamma}(y_{n+1}|\boldsymbol{x})\\&=p^n_{\mathscr{X},\gamma}(\boldsymbol{x})P^{n+1}_{\mathscr{X},\gamma}(P^{n+2}_{\mathscr{X},\gamma}(f|\boldsymbol{x},\cdot)|\boldsymbol{x}).\end{aligned}$$

Since the linear prevision $P^N_{\mathscr{X},\gamma}$ is exchangeable, we see that $P^N_{\mathscr{X},\gamma}(g)=P^N_{\mathscr{X},\gamma}(g')$. Hence $P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x})=P^{n+1}_{\mathscr{X},\gamma}(P^{n+2}_{\mathscr{X},\gamma}(f|\boldsymbol{x},\cdot)|\boldsymbol{x})$, since $p^n_{\mathscr{X},\gamma}(\boldsymbol{x})>0$, by the assumption of regular exchangeability. Taking the infimum on both sides over all γ in Γ, and invoking regular exchangeability leads to

$$\begin{aligned}\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})&=\inf_{\gamma\in\Gamma}P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x})=\inf_{\gamma\in\Gamma}P^{n+1}_{\mathscr{X},\gamma}(P^{n+2}_{\mathscr{X},\gamma}(f|\boldsymbol{x},\cdot)|\boldsymbol{x})\\&\geq\inf_{\gamma\in\Gamma}P^{n+1}_{\mathscr{X},\gamma}(\inf_{\gamma'\in\Gamma}P^{n+2}_{\mathscr{X},\gamma'}(f|\boldsymbol{x},\cdot)|\boldsymbol{x})=\underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{x},\cdot)|\boldsymbol{x}).\end{aligned}$$

Now recall that $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})=\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ and $\underline{P}^{n+2}_{\mathscr{X}}(\cdot|\boldsymbol{x},z)=\underline{P}^{n+2}_{\mathscr{X}}(\cdot|\boldsymbol{m}+\boldsymbol{e}_z)$. □

Proof of Proposition 3 Statement 1 follows from (separate) coherence [use (C1) and (C4)]. To prove statement 2, fix $0\leq n\leq N-1$ and non-negative k and ℓ such that $k+\ell\leq n$. Consider a set $\mathscr{X}$ with three elements a, b and c, then there is always an $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ such that $m_a=k$ and $m_b=\ell$ (whence $m_c=n-k-\ell\geq 0$). Consider the proper subsets $A=\{a\}$ and $B=\{b\}$ of $\mathscr{X}$, then their union $A\cup B=\{a,b\}$ is a proper subset of $\mathscr{X}$ and their intersection is empty: $A\cap B=\emptyset$. Now use

the super-additivity of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ [this follows from (C2)] and then representation insensitivity to find that indeed

$$\varphi(n,k+\ell)=\underline{P}^{n+1}_{\mathscr{X}}(A\cup B|\boldsymbol{m})\geq\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m})+\underline{P}^{n+1}_{\mathscr{X}}(B|\boldsymbol{m})=\varphi(n,k)+\varphi(n,\ell).$$

Statements 3–5 follow trivially from statements 1 and 2. To prove statement 6, consider a set of categories $\mathscr{X}=\{a,b\}$. Fix $0\leq k\leq n\leq N-2$, and let $\boldsymbol{m}\in\mathscr{N}^n_{\mathscr{X}}$ be such that $m_a=k$ and $m_b=n-k$. We apply Proposition 2 with $f=I_{\{a\}}$ to get $\underline{P}^{n+1}_{\mathscr{X}}(\{a\}|\boldsymbol{m})\geq\underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(\{a\}|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m})$. Now define the gamble g on $\mathscr{X}$ by $g(a):=\underline{P}^{n+2}_{\mathscr{X}}(\{a\}|\boldsymbol{m}+\boldsymbol{e}_a)=\varphi(n+1,k+1)$ and $g(b):=\underline{P}^{n+2}_{\mathscr{X}}(\{a\}|\boldsymbol{m}+\boldsymbol{e}_b)=\varphi(n+1,k)$, then it is clear from statement 5 that $g(a)\geq g(b)$ and therefore, using $g=g(b)+[g(a)-g(b)]I_{\{a\}}$ and the (separate) coherence of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ [use (C3) and (C6)], $\underline{P}^{n+1}_{\mathscr{X}}(\{a\}|\boldsymbol{m})\geq\underline{P}^{n+1}_{\mathscr{X}}(g|\boldsymbol{m})=g(b)+[g(a)-g(b)]\underline{P}^{n+1}_{\mathscr{X}}(\{a\}|\boldsymbol{m})$. If we now recall that $\underline{P}^{n+1}_{\mathscr{X}}(\{a\}|\boldsymbol{m})=\varphi(n,k)$, we are done. Let us prove statement 7. Observe that for $0\leq k\leq n\leq N-2$, $\varphi(n,k)\geq 0$ and $\varphi(n+1,k+1)\geq\varphi(n+1,k)$ [by statement 5]. Statement 6 then implies that indeed $\varphi(n,k)\geq\varphi(n+1,k)$. To prove 8, apply statement 6 with $1=k\leq n\leq N-2$ to find that

$$\begin{aligned}\varphi(n,1)&\geq\varphi(n+1,1)+\varphi(n,1)[\varphi(n+1,2)-\varphi(n+1,1)]\\&\geq\varphi(n+1,1)+\varphi(n,1)[2\varphi(n+1,1)-\varphi(n+1,1)]\\&=\varphi(n+1,1)[1+\varphi(n,1)],\end{aligned}$$

where the second inequality follows from statement 4. We turn to statement 9. Observe that $1\geq\varphi(n,n)\geq n\varphi(n,1)$ by statement 4, whence $\frac{1}{\varphi(n,1)}\geq n$ and therefore indeed $s_n\geq 0$. To prove that s_n is non-decreasing, apply the inequality in statement 8 for $1\leq n\leq N-2$ to get, after division of both sides of the inequality by $\varphi(n+1,1)\varphi(n,1)$: $s_{n+1}+n+1=1/\varphi(n+1,1)\geq 1/\varphi(n,1)+1=s_n+n+1$. □

Proof of Proposition 5 Consider any gamble f on $\mathscr{X}$, then we have to prove that $\underline{P}(f)=\inf f$. Since, as a consequence of the coherence of $\underline{P}$ [use (C6)], $\underline{P}(f)=\inf f+\underline{P}(f-\inf f)$, we only need to prove that $\underline{P}(g)=0$, where g is any non-negative gamble with $\inf g=0$. For any positive integer n, the set $A_n:=\{g>\frac{1}{n}\}$ is different from $\mathscr{X}$ because $\inf g=0$, so the assumption implies that $\underline{P}(A_n)=0$. Since moreover $g\leq\frac{1}{n}+I_{A_n}\sup g$, we deduce from the coherence of $\underline{P}$ [use (C5), (C3) and (C6)] that $0\leq\underline{P}(g)\leq\frac{1}{n}+\underline{P}(A_n)\sup g=\frac{1}{n}$ for all n, whence $\underline{P}(g)=0$. □

Proof of Theorem 6 We first prove that ν^N is regularly exchangeable. Consider the collection Γ of all maps that associate with any non-empty and finite set $\mathscr{X}$, some element $\gamma(\mathscr{X})$ of $\{\boldsymbol{\alpha}\in\mathbb{R}_+^{\mathscr{X}}:\sum_{z\in\mathscr{X}}\alpha_z=1\}$, where $\mathbb{R}_+$ is the set of (strictly) positive real numbers. For each γ in Γ, consider the predictive system σ^N_γ of predictive linear previsions $P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x}):=\sum_{z\in\mathscr{X}}\gamma_z(\mathscr{X})f(z)$, with predictive mass functions $p^{n+1}_{\mathscr{X},\gamma}(z|\boldsymbol{x}):=\gamma_z(\mathscr{X})>0$, $z\in\mathscr{X}$. Then it is clear that for all $\boldsymbol{x}$ in $\mathscr{X}^{N-1}$,

$p^{N-1}_{\mathscr{X},\gamma}(\boldsymbol{x}) = \prod_{k=0}^{N-2} p^{k+1}_{\mathscr{X},\gamma}(x_{k+1}|x_1,\ldots,x_k) > 0$, and that the vacuous predictive system is the infimum of the collection σ^N_γ, $\gamma \in \Gamma$. The corresponding joint mass functions $p^N_{\mathscr{X},\gamma}$ are given by $p^N_{\mathscr{X},\gamma}(\boldsymbol{x}) = \prod_{z\in\mathscr{X}} \gamma_z(\mathscr{X})^{T_z(\boldsymbol{x})}$, $\boldsymbol{x} \in \mathscr{X}^N$. As these only depend on $\boldsymbol{x}$ through $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$, the precise predictive systems σ^N_γ are exchangeable. Therefore all conditions for regular exchangeability are satisfied. That ν^N is representation insensitive, follows immediately from

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \min_{x\in\mathscr{X}} f(x) = \min_{r\in f(\mathscr{X})} r = \min_{r\in f(\mathscr{X})} \mathrm{id}_{f(\mathscr{X})}(r) = \underline{P}^{n+1}_{f(\mathscr{X})}(\mathrm{id}_{f(\mathscr{X})}|\boldsymbol{m}^f).$$

The lower probability function for ν^N satisfies $\varphi(n,m) = \min_{r\in\{0,1\}} \mathrm{id}_{\{0,1\}}(r) = 0$, for $0 \leq m \leq n \leq N-1$. Finally, since the vacuous lower prevision is point-wise dominated by all linear previsions, the predictive system ν^N, which consists only of vacuous lower previsions, is the point-wise smallest coherent predictive system. We deduce that it is the bottom of the structure $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$. $\square$

Proof of Theorem 7 Consider any finite and non-empty set of categories $\mathscr{X}$, and let $0 < n \leq N-1$ and $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$. It follows from representation insensitivity and the linearity of $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ that for any gamble f on $\mathscr{X}$

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \sum_{z\in\mathscr{X}} f(z) P^{n+1}_{\mathscr{X}}(\{z\}|\boldsymbol{m}) = \sum_{z\in\mathscr{X}} f(z)\varphi(n,m_z), \qquad (\star)$$

so, taking f to be the constant function 1, it follows that $\sum_{z\in\mathscr{X}} \varphi(n,m_z) = 1$. Using another consequence of representation insensitivity and linearity [Corollary 4], we infer that $\varphi(n,m_z) = m_z\varphi(n,1)$, so $1 = \sum_{z\in\mathscr{X}} m_z\varphi(n,1) = \varphi(n,1)\sum_{z\in\mathscr{X}} m_z = n\varphi(n,1)$. We see that $\varphi(n,1) = \frac{1}{n}$ for $n > 0$ and Corollary 4 then implies $\varphi(n,k) = \frac{k}{n}$. Substituting this back into Eq. ($\star$) yields Eq. (5).

We still have to show that π^N is exchangeable and representation insensitive. We begin with exchangeability, and establish that π^N is the lower envelope of a specific collection $\sigma^N_\gamma, \gamma \in \Gamma$ of exchangeable systems of predictive linear previsions. Consider the collection Γ of all maps γ that associate with any finite and non-empty set $\mathscr{X}$, some particular element $\gamma(\mathscr{X})$ of $\mathscr{X}$. Now define the predictive system σ^N_γ as follows: the predictive linear previsions are given by $P^1_{\mathscr{X},\gamma}(f) = f(\gamma(\mathscr{X}))$ for any $f \in \mathscr{L}(\mathscr{X})$, and by $P^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{m}) = S^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ [Eq. (5)] for $0 < n \leq N-1$. The resulting joint linear prevision $P^N_{\mathscr{X},\gamma}$ has a joint mass function determined by $p^N_{\mathscr{X},\gamma}(\gamma(\mathscr{X}),\ldots,\gamma(\mathscr{X})) = 1$. It is permutation invariant, and therefore the predictive system σ^N_γ is exchangeable. It is straightforward to check that π^N is indeed the lower envelope of the collection of exchangeable precise predictive systems σ^N_γ, $\gamma \in \Gamma$, and is therefore an exchangeable predictive system. To check that it is rep-

resentation insensitive, observe that for any gamble f on $\mathscr{X}$ and all $0 < n \leq N-1$:

$$S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \sum_{z\in\mathscr{X}} f(x)\frac{m_z}{n} = \sum_{r\in f(\mathscr{X})} r\frac{\sum_{f(x)=r} m_x}{n} = S^{n+1}_{f(\mathscr{X})}(\mathrm{id}_{f(\mathscr{X})}|\boldsymbol{m}^f). \qquad \square$$

Proof of Theorem 9 Consider the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ that belong to the predictive system σ^N. For any non-negative gamble g on $\mathscr{X}$, it follows from the (separate) coherence [use (C2) and (C6)] of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ and representation insensitivity that

$$\underline{P}^{n+1}_{\mathscr{X}}(g|\boldsymbol{m}) \geq \sum_{z\in\mathscr{X}} g(z)\underline{P}^{n+1}_{\mathscr{X}}(\{z\}|\boldsymbol{m}) = \sum_{z\in\mathscr{X}} g(z)\varphi(n,m_z)$$

and if we use Proposition 3.4 and the assumption, we get

$$\underline{P}^{n+1}_{\mathscr{X}}(g|\boldsymbol{m}) \geq \sum_{z\in\mathscr{X}} g(z)m_z\varphi(n,1) \geq \varepsilon_n \sum_{z\in\mathscr{X}} g(z)\frac{m_z}{n} = \varepsilon_n S^{n+1}_{\mathscr{X}}(g|\boldsymbol{m}).$$

Again using the (separate) coherence [(C6)] of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$, it follows that for any gamble f on $\mathscr{X}$, since $f - \min f$ is non-negative,

$$\begin{aligned}\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) &= \min f + \underline{P}^{n+1}_{\mathscr{X}}(f-\min f|\boldsymbol{m}) \\ &\geq \min f + \varepsilon_n S^{n+1}_{\mathscr{X}}(f-\min f|\boldsymbol{m}) \\ &= \varepsilon_n S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + (1-\varepsilon_n)\min f.\end{aligned}$$

If we compare this with Eq. (6), we see that $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ point-wise dominates the corresponding predictive lower prevision in σ^N_ε, whence indeed $\sigma^N_\varepsilon \preceq \sigma^N$. $\square$

Proof of Theorem 10 Consider the IDMM-predictive system defined by fixing some $s > 0$ in Eq. (10). From Section 5.2, it only remains to prove that it is regularly exchangeable. Consider the collection Γ of all maps that associate with any non-empty and finite set $\mathscr{X}$, some element $\gamma(\mathscr{X})$ of $\left\{\boldsymbol{\alpha} \in \mathbb{R}^{\mathscr{X}}_+ : \sum_{z\in\mathscr{X}} \alpha_z = 1\right\}$. For each γ in Γ, consider the system σ^N_γ of predictive linear previsions

$$P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x}) = \frac{n}{n+s}S_{\mathscr{X}}(f|\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})) + \frac{s}{n+s}\sum_{z\in\mathscr{X}} \gamma_z(\mathscr{X})f(z),$$

with predictive mass functions $p^{n+1}_{\mathscr{X},\gamma}(z|\boldsymbol{x}) = \frac{T_z(\boldsymbol{x})+s\gamma_z(\mathscr{X})}{n+s} > 0$, $z \in \mathscr{X}$. Then it is clear that for all $\boldsymbol{x}$ in $\mathscr{X}^{N-1}$, $p^{N-1}_{\mathscr{X},\gamma}(\boldsymbol{x}) = \prod_{k=0}^{N-2} p^{k+1}_{\mathscr{X},\gamma}(x_{k+1}|x_1,\ldots,x_k) > 0$, and that the IDMM-predictive system is the infimum of the collection σ^N_γ, $\gamma \in \Gamma$. It is readily checked that the corresponding joint mass functions $p^N_{\mathscr{X},\gamma}$ are given by

$$p^N_{\mathscr{X},\gamma}(\boldsymbol{x}) = \frac{1}{\binom{N+s-1}{N}}\prod_{z\in\mathscr{X}}\binom{T_z(\boldsymbol{x})+s\gamma_z(\mathscr{X})-1}{T_z(\boldsymbol{x})},$$

where $\binom{r}{k} = \frac{1}{k!}\prod_{i=0}^{k-1}(r-i)$ for real r and $k > 0$, and $\binom{r}{0} = 1$. As these only depend on $\boldsymbol{x}$ through $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$, the precise predictive systems σ_γ^N are exchangeable. Therefore all conditions for regular exchangeability are satisfied. □

Proof of Theorem 11 We write down the left-hand side of Eq. (11) using Eq. (6) and $\varepsilon_n = n/(n+s_n) > 0$ [see Proposition 8]. Since A is a proper subset of $\mathscr{X}$, this results in

$$\begin{aligned}
\underline{P}_{\mathscr{X}}^{n+1}(I_A[f-\mu]) &= \frac{n}{n+s_n}\sum_{x\in A}[f(x)-\mu]\frac{m_x}{n} + \frac{s_n}{n+s_n}\min\{0, \min_{x\in A} f(x)-\mu\} \\
&= \frac{m_A}{n+s_n}\left[\sum_{x\in A} f(x)\frac{m_x}{m_A} - \mu\right] + \frac{s_n}{n+s_n}\min\{0, \min_{x\in A} f(x)-\mu\} \\
&= \frac{m_A}{n+s_n}[S_A^{m_A+1}(f_A|\boldsymbol{m}_A)-\mu] + \frac{s_n}{n+s_n}\min\{0, \min f_A-\mu\}.
\end{aligned}$$

This value can only be zero if $\mu \geq \min f_A$, so we see that Eq. (11) is equivalent to

$$\mu = \underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A) = \frac{m_A}{m_A+s_n}S_A^{m_A+1}(f_A|\boldsymbol{m}_A) + \frac{s_n}{m_A+s_n}\min f_A.$$

Comparing this to $\underline{P}_A^{m_A+1}(f_A|\boldsymbol{m}_A) = \frac{m_A}{m_A+s_{m_A}}S_A^{m_A+1}(f_A|\boldsymbol{m}_A) + \frac{s_{m_A}}{m_A+s_{m_A}}\min f_A$, we see that $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A)$ is equal to $\underline{P}_A^{m_A+1}(f_A|\boldsymbol{m}_A)$ if and only if

$$\frac{m_A(s_n-s_{m_A})}{(m_A+s_n)(m_A+s_{m_A})}\left[S_A^{m_A+1}(f_A|\boldsymbol{m}_A) - \min f_A\right] = 0.$$

We want this equality to hold for all gambles f on all $\mathscr{X}$, all $n = 1,\ldots,N-1$, all $\boldsymbol{m} \in \mathscr{N}_{\mathscr{X}}^n$, and all proper subsets A of $\mathscr{X}$ such that $m_A > 0$. It is clear that the condition $s_n = s$ for some $s > 0$ and all $n = 1,\ldots,N-1$ is sufficient. To show that it is also necessary, fix $n \in \{2,\ldots,N-1\}$ and choose $\mathscr{X} = \{a,b,c\}$, $A = \{a,b\}$, a gamble f on $\mathscr{X}$ such that $f(a) > f(b) = 0$, and $\boldsymbol{m} \in \mathscr{N}_{\mathscr{X}}^n$ such that $m_a = n-1$, $m_b = 0$ and $m_c = 1$. Then the condition above becomes $\frac{(n-1)(s_n-s_{n-1})}{(s_n+n-1)(s_{n-1}+n-1)}f(a) = 0$, or in other words $s_n = s_{n-1}$. □

References

[1] P.-S. Laplace, Philosophical Essay on Probabilities, Dover Publications, 1951, English translation of [20].

[2] R. Carnap, The continuum of inductive methods, The University of Chicago Press, 1952.

[3] P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society, Series B 58 (1996) 3–57, with discussion.

[4] P. Walley, J.-M. Bernard, Imprecise probabilistic prediction for categorical data, Tech. Rep. CAF-9901, Laboratoire Cognition et Activitées Finalisées, Université de Paris 8 (January 1999).

[5] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.

[6] P. Walley, Measures of uncertainty in expert systems, Artificial Intelligence 83 (1996) 1–58.

[7] G. de Cooman, E. Miranda, Symmetry of models versus models of symmetry, in: W. L. Harper, G. R. Wheeler (Eds.), Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr., King's College Publications, 2006, accepted for publication.

[8] G. de Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Artificial Intelligence 159 (2004) 75–125.

[9] B. de Finetti, Theory of Probability, John Wiley & Sons, Chichester, 1974–1975, English translation of [21], two volumes.

[10] P. M. Williams, Notes on conditional previsions, International Journal of Approximate Reasoning 44 (2007) 366–383, revised journal version of [22].

[11] E. Miranda, G. de Cooman, Marginal extension in the theory of coherent lower previsions, International Journal of Approximate Reasoning. In press.

[12] B. de Finetti, La prévision: ses lois logiques, ses sources subjectives, Annales de l'Institut Henri Poincaré 7 (1937) 1–68, English translation in [23].

[13] B. A. Davey, H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, Cambridge, 1990.

[14] G. de Cooman, M. C. M. Troffaes, E. Miranda, n-Monotone lower previsions, Journal of Intelligent and Fuzzy Systems 16 (2005) 253–263.

[15] G. de Cooman, M. C. M. Troffaes, E. Miranda, n-Monotone exact functionals. Submitted for publication.

[16] W. E. Johnson, Logic, Part III. The Logical Foundations of Science, Cambridge University Press, 1924, reprinted by Dover Publications in 1964.

[17] S. L. Zabell, W. E. Johnson's "sufficientness" postulate, The Annals of Statistics 10 (1982) 1090–1099, reprinted in [24].

[18] H. Jeffreys, Theory of Probability, Oxford Classics series, Oxford University Press, 1998, reprint of the third edition (1961), with corrections.

[19] J.-M. Bernard, Bayesian analysis of tree-structured categorized data, Revue Internationale de Systémique 11 (1997) 11–29.

[20] P.-S. Laplace, Essai philosophique sur les probabilités, Christian Bourgois Éditeur, 1986, reprinted from the fifth edition (1825).

[21] B. de Finetti, Teoria delle Probabilità, Einaudi, Turin, 1970.

[22] P. M. Williams, Notes on conditional previsions, Tech. rep., School of Mathematical and Physical Science, University of Sussex, UK (1975).

[23] H. E. Kyburg Jr., H. E. Smokler (Eds.), Studies in Subjective Probability, Wiley, New York, 1964, second edition (with new material) 1980.

[24] S. L. Zabell, Symmetry and Its Discontents: Essays on the History of Inductive Probability, Cambridge Studies in Probability, Induction, and Decision Theory, Cambridge University Press, Cambridge, UK, 2005.