

Immediate prediction under exchangeability & representation insensitivity

Gert de Cooman, Enrique Miranda & Erik Quaeghebeur

1 The setting

Sampling: A subject makes a fixed number $N > 0$ of successive observations, represented by *random variables* X_1, \dots, X_N . For example, when drawing coloured balls without replacement from an urn, X_k designates the unknown colour of the k -th ball.

Immediate prediction: The subject in some way uses zero or more observations X_1, \dots, X_n made previously (so n belongs to $\{0, 1, \dots, N-1\}$), to predict, or make inferences about, the value of the *next* observation X_{n+1} .

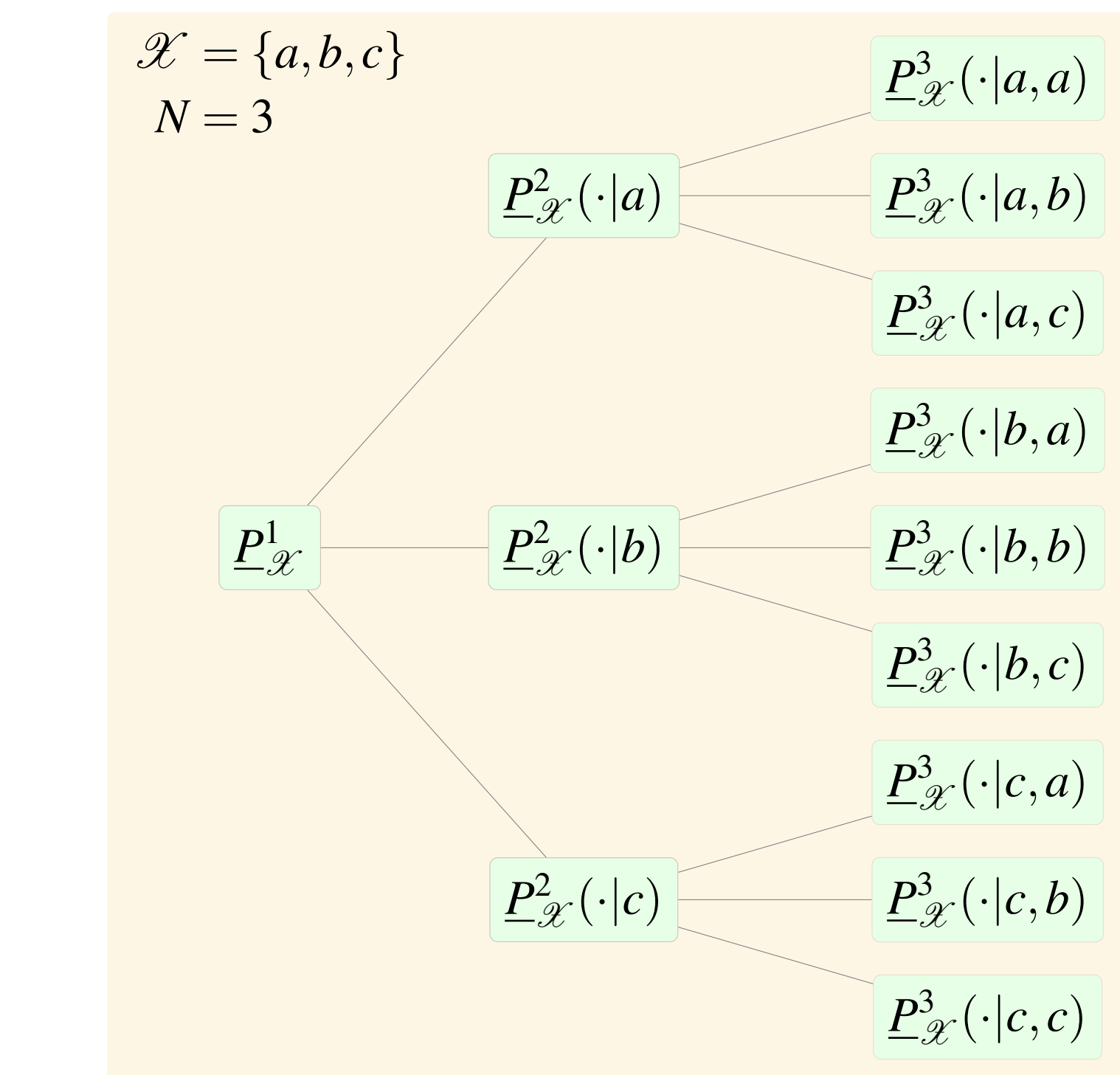
Families of predictive lower previsions: The subject can determine, beforehand, a finite and non-empty set \mathcal{X} of possible values, or *categories*, for the random variables.

For each n and each sequence $\mathbf{x} = (x_1, \dots, x_n)$ in \mathcal{X}^n , she can give a *predictive lower prevision* $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ for X_{n+1} , given the values $(X_1, \dots, X_n) = (x_1, \dots, x_n) = \mathbf{x}$ of the previous observations. It is defined on the set of all gambles f on \mathcal{X} .

$\mathcal{X} = \{a, b, c\}$ Let $f(a) = 1, f(b) = 3, f(c) = -2$,
 $N = 3$ then, e.g., $P_{\mathcal{X}}^3(f|c, a) = -\frac{1}{2}$.
 $n = 2$ Let $A = \{a, c\}$,
 $\mathbf{x} = (c, a)$ then, e.g., $P_{\mathcal{X}}^3(A|c, a) = \frac{4}{5}$.

An \mathcal{X} -family $\sigma_{\mathcal{X}}^N$ of predictive lower previsions is the set formed for all possible observations:

$$\sigma_{\mathcal{X}}^N := \{P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n \text{ and } n = 0, 1, \dots, N-1\}.$$



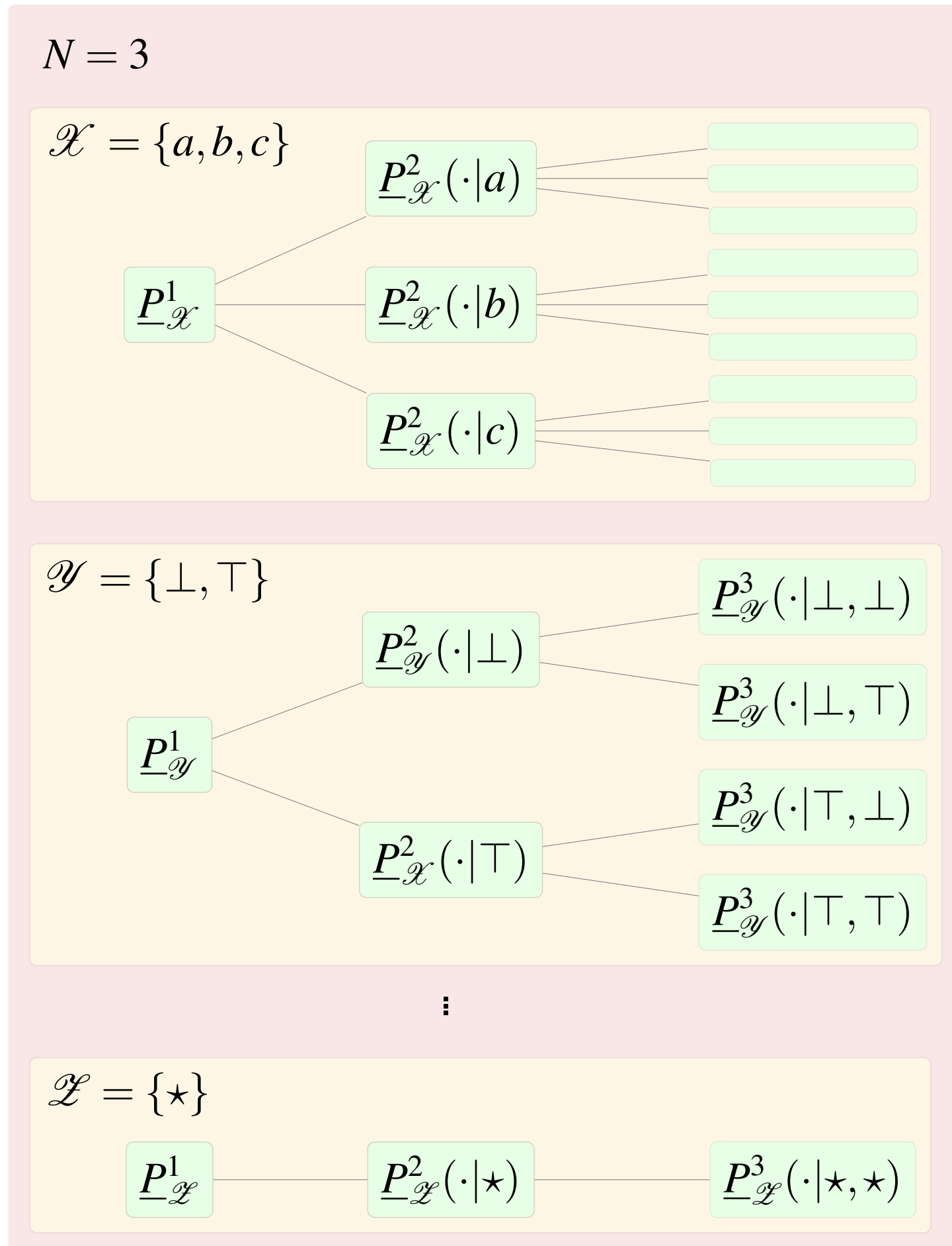
Precise predictive families are those that only contain predictive *linear* previsions $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$. With each of these, there corresponds a predictive probability mass function. They in turn allow us, using Bayes's rule, to find the unique joint probability mass functions $p_{\mathcal{X}}^n$ on \mathcal{X}^n and the corresponding *joint linear prevision* $P_{\mathcal{X}}^n$, which models beliefs about the values that the random variables (X_1, \dots, X_N) assume *jointly* in \mathcal{X}^N .

Acknowledgements

Jean-Marc Bernard, Frank-Coolen, Thomas Augustin, and the reviewers.

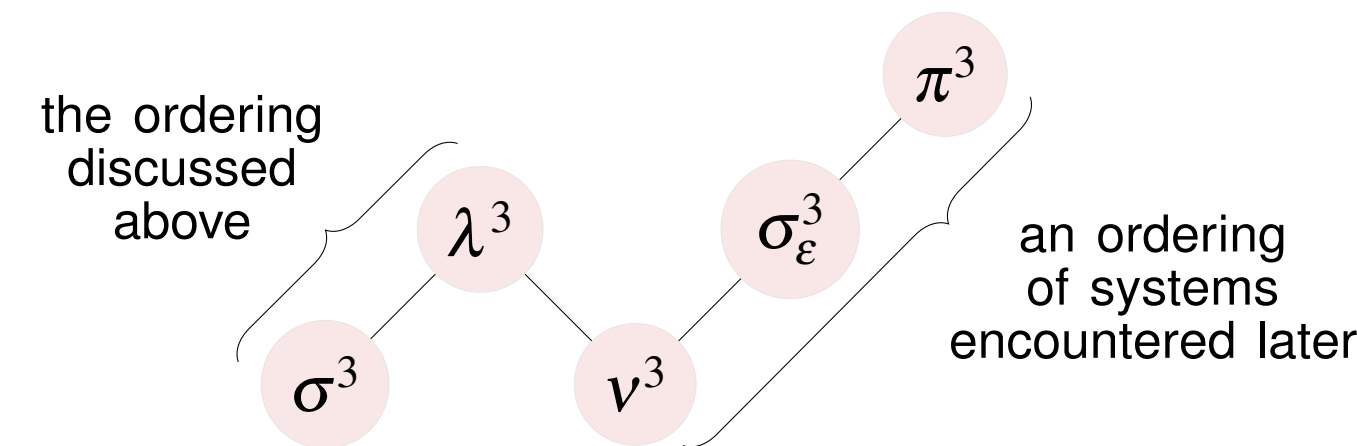
Systems of predictive lower previsions: The inferences or predictions of a predictive \mathcal{X} -family might depend on the actual choice of \mathcal{X} made. So we let our subject consider predictive families for all conceivable choices of \mathcal{X} . We collect these families in a *system* σ^N of predictive lower previsions:

$$\sigma^N := \{P_{\mathcal{X}}^N : \mathcal{X} \text{ is a finite and non-empty set}\}.$$



Precise predictive systems are those that only contain precise predictive families.

Predictive systems can be *partially ordered*: The system σ^N is *more conservative* than the system λ^N , if each predictive lower prevision $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ in σ^N is point-wise dominated by the corresponding predictive lower prevision $Q_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ in λ^N .



A collection $\{\sigma_{\gamma}^N : \gamma \in \Gamma\}$ of predictive systems may have an infimum with respect to this partial order. Whenever it exists, this infimum system σ^N can be seen as a *lower envelope*: each of its predictive lower previsions $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ is defined as the lower envelope $\inf_{\gamma \in \Gamma} P_{\mathcal{X}}^{n+1, \gamma}(\cdot|\mathbf{x})$ of the predictive lower previsions in the predictive systems σ_{γ}^N .

Selected references

- G. de Cooman and E. Miranda. Symmetry of models versus models of symmetry. In W. L. Harper and G. R. Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*, pages 67–149. 2007.
- P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Université de Paris 8, 1999.
- S. L. Zabell. W. E. Johnson's "sufficiency" postulate. *The Annals of Statistics*, 10:1090–1099, 1982.

2 Requirements & Assumptions

Coherence: Coherence is a requirement on the individual predictive lower previsions.

A predictive system is called *coherent* if it is the lower envelope of a collection of precise predictive systems. This is equivalent to requiring that all the predictive lower previsions $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ in the system should be separately coherent.

(Regular) exchangeability: Exchangeability is an assumption about a family of predictive lower previsions.

A precise predictive system is *exchangeable* if all the associated joint linear previsions $P_{\mathcal{X}}^N$ are exchangeable, i.e., invariant under permutation of the random variables X_1, \dots, X_N .

A general predictive system is called *exchangeable* if it is the lower envelope of a collection $\{\sigma_{\gamma}^N : \gamma \in \Gamma\}$ of exchangeable precise predictive systems. It is *regularly exchangeable* if all predictive linear previsions $P_{\mathcal{X}, \gamma}^{n+1}(\cdot|\mathbf{x})$ in each of these systems σ_{γ}^N can be uniquely derived from the joint linear prevision $P_{\mathcal{X}, \gamma}^N$ by applying Bayes's rule. (For this, the joint mass functions $p_{\mathcal{X}, \gamma}^n$ should be strictly positive for $n < N$.)

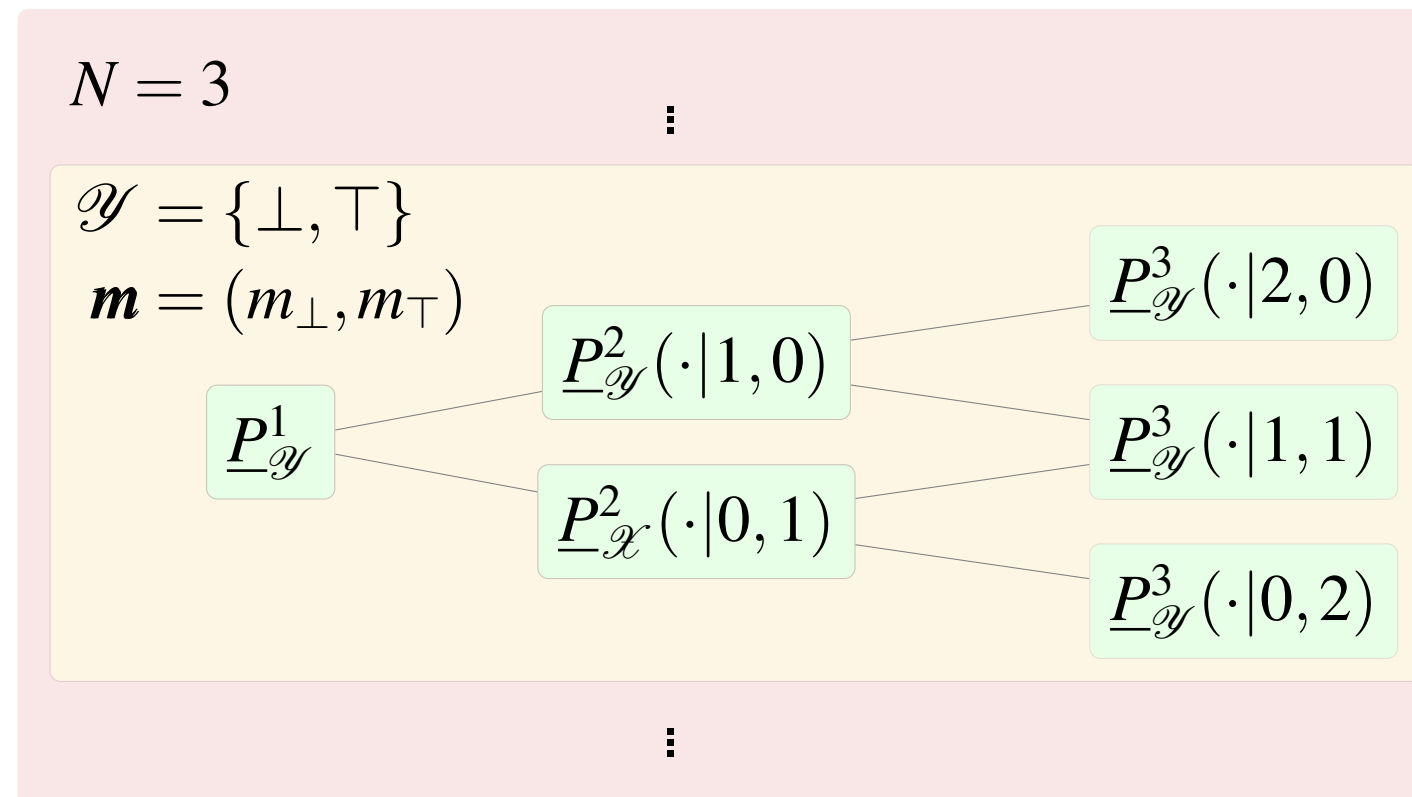
3 Some results

From sequences of observations to count vectors: In any regularly exchangeable predictive system, the predictive lower previsions $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{x})$ only depend on the sequence of observations \mathbf{x} through its *count vector* $\mathbf{m} \in \mathcal{N}_{\mathcal{X}}^n$, with

$$m_z := |\{k \in \{1, \dots, n\} : x_k = z\}|,$$

$$\mathcal{N}_{\mathcal{X}}^n := \{\mathbf{m} \in \mathbb{N}_0^n : \sum_{z \in \mathcal{X}} m_z = n\}.$$

All predictive lower previsions for given sequences with the same count vector \mathbf{m} can therefore be written as $P_{\mathcal{X}}^{n+1}(\cdot|\mathbf{m})$.



So, for regularly exchangeable predictive systems, count vectors are a *sufficient statistic*. From now on, we only consider (possibly non-exchangeable) predictive systems for which this is the case.

A useful (in)equality In any regularly exchangeable predictive system, it holds for all gambles f that

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) \geq P_{\mathcal{X}}^{n+1}(P_{\mathcal{X}}^{n+2}(f|\mathbf{m} + \mathbf{e}_x)|\mathbf{m}),$$

where $n \leq N-2$ and $\mathbf{e}_x \in \mathcal{N}_{\mathcal{X}}^1$ for $x \in \mathcal{X}$ such that, using the Kronecker delta, $(\mathbf{e}_x)_z = \delta_{xz}$. For precise regularly exchangeable predictive systems, this becomes a 'useful equality'.

4 Some more requirements

Representation insensitivity: Representation insensitivity is a requirement that works between predictive lower previsions for the same number of observations.

It comprises three invariance requirements:

- pooling invariance:** inferences that do not depend on the distinction between some categories should stay the same when those categories are pooled;
- renaming invariance:** apart from avoiding confusion, the names of the categories should not matter;
- category permutation invariance:** in a state of prior ignorance, which we consider here, the subject has no reason to distinguish between the categories, so the inferences should be invariant under a permutation of them.

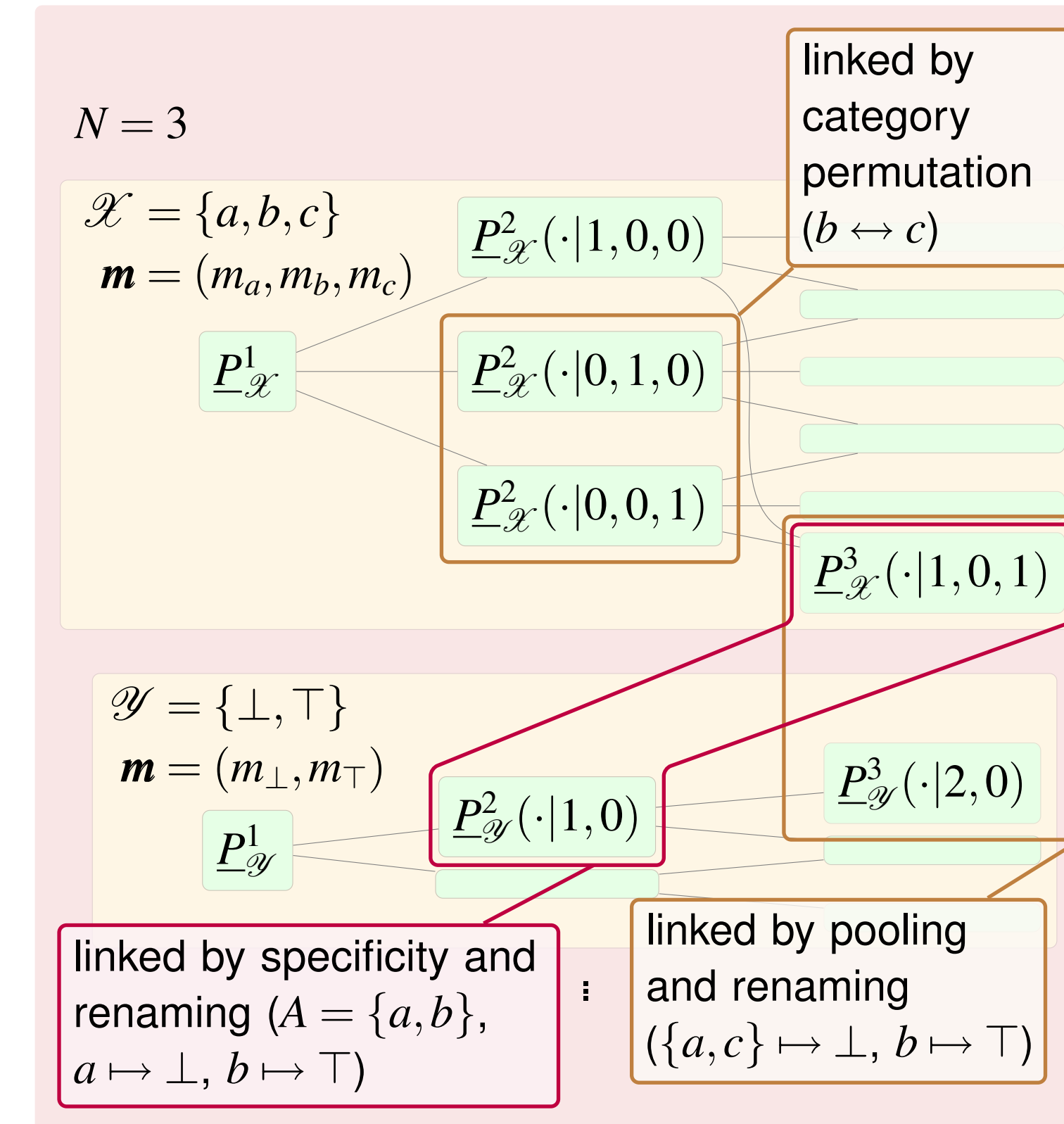
Combining these, we can say a predictive system is *representation insensitive* if for all n , for any category sets \mathcal{X} and \mathcal{Y} , for any $\mathbf{m} \in \mathcal{N}_{\mathcal{X}}^n$ and $\mathbf{m}' \in \mathcal{N}_{\mathcal{Y}}^n$, and for any gambles f on \mathcal{X} and g on \mathcal{Y} with identical ranges, the following holds:

$$\mathbf{m}^f = \mathbf{m}'^g \Rightarrow P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) = P_{\mathcal{Y}}^{n+1}(g|\mathbf{m}'),$$

with $\mathbf{m}^f := \sum_{f(x)=r} m_x$. This means $P_{\mathcal{X}}^{n+1}(f|\mathbf{m})$ only depends on the values that f may assume, and on the number of times each value has been observed:

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) = P_{f(\mathcal{X})}^{n+1}(\text{id}_{f(\mathcal{X})}|\mathbf{m}^f),$$

where $\text{id}_{f(\mathcal{X})}$ is the identity map on the range of f .



Specificity (optional): Specificity is a requirement that works between predictive lower previsions for a different number of observations related by pooling.

An exchangeable predictive system is *specific* if for all gambles f and all non-trivial events A in \mathcal{X} containing a non-zero number m_A of observations, it holds that

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}, A) = P_A^{n+1}(f_A|\mathbf{m}_A),$$

where f_A and \mathbf{m}_A are the restriction of f and \mathbf{m} to A . So, knowing that the $(n+1)$ -th observation belongs to A allows you to ignore all the previous observations that lie outside A .

5 More results

The lower probability function: With any predictive system we associate a map φ defined for all n and $k \leq n$ by

$$\varphi(n, k) := P_{\{0,1\}}^{n+1}(\text{id}_{\{0,1\}}|n-k, k).$$

For representation insensitive systems it fully characterizes all predictive lower probabilities (cf. *Johnson's sufficientness postulate*) and is therefore called the *lower probability function*; to wit, let A be some event, and m_A the associated number of observations, then

$$P_{\mathcal{X}}^{n+1}(A|\mathbf{m}) = P_{\{0,1\}}^{n+1}(\text{id}_{\{0,1\}}|n-m_A, m_A) = \varphi(n, m_A).$$

It allows us to draw intuitively appealing conclusions, which are valid in any coherent representation insensitive system:

- the lower/upper probability of observing an event that has not/always been observed before is zero/one;
- if n remains fixed, then both the lower and upper probability of observing A again do not decrease if m_A increases;
- in systems that are also regularly exchangeable: if m_A remains the same as n increases, then the lower probability for observing A again does not increase.

Some representation insensitive exchangeable systems: To start: all the $P_{\mathcal{X}}^1$ in a representation insensitive and exchangeable predictive system must be vacuous.

A subject that is too conservative to learn uses the regularly exchangeable *vacuous predictive system* v^N . All its predictive lower previsions are vacuous, so $P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) := \min f$.

A subject that believes that categories unobserved in the past remain so in the future, uses the (not regularly) exchangeable *Haldane predictive system* π^N . For $n > 0$, all its predictive previsions are linear and strongly tied to the observations:

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) = S_{\mathcal{X}}^{n+1}(f|\mathbf{m}) := \sum_{z \in \mathcal{X}} f(z) \frac{m_z}{n}.$$

Other systems can be formed as convex mixtures of the two extreme ones above. We define *mixing predictive systems* σ_{ε}^N with a $[0, 1]$ -bounded *mixing sequence* ε of length N and

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) := \varepsilon_n S_{\mathcal{X}}^{n+1}(f|\mathbf{m}) + (1 - \varepsilon_n) \min f;$$

note that implicitly $\varepsilon_0 = 0$. Representation insensitivity is retained after mixing; a sufficient condition for regular exchangeability is the reformulated 'useful inequality' $\frac{\varepsilon_n}{n} \geq \frac{\varepsilon_{n+1}}{n+1} (1 + \frac{\varepsilon_n}{n})$.

The lower probability function of a mixing system is given by $\varphi(n, k) = \varepsilon_n \frac{k}{n}$. So, as $\varepsilon_n = n\varphi(n, 1)$, a mixing system can be defined by specifying the lower probability of observing any non-trivial event that has been observed once in n trials; it is the most conservative system with these lower probabilities.

The imprecise Dirichlet-Multinomial model: Any mixing system that is specific or for which the 'useful equality' holds, is uniquely characterized by some $s > 0$ such that $\varepsilon_n = \frac{n}{n+s}$ and thus

$$P_{\mathcal{X}}^{n+1}(f|\mathbf{m}) = \frac{n}{n+s} S_{\mathcal{X}}^{n+1}(f|\mathbf{m}) + \frac{s}{n+s} \min f.$$

This regularly exchangeable representation insensitive predictive system is related to the imprecise Dirichlet-Multinomial model with hyper-parameter s .