Valid inferential models for prediction in supervised learning problems[*]

Leonardo Cella[†] and Ryan Martin[†]

June 13, 2022

Abstract

Prediction, where observed data is used to quantify uncertainty about a future observation, is a fundamental problem in statistics. Prediction sets with coverage probability guarantees are a common solution, but these do not provide probabilistic uncertainty quantification in the sense of assigning beliefs to relevant assertions about the future observable. Alternatively, we recommend the use of a *probabilistic predictor*, a data-dependent (imprecise) probability distribution for the to-be-predicted observation given the observed data. It is essential that the probabilistic predictor be reliable or valid, and here we offer a notion of validity and explore its behavioral and statistical implications. In particular, we show that valid probabilistic predictors must be imprecise, that they avoid sure loss, and that they lead to prediction procedures with desirable frequentist error rate control properties. We provide a general construction of a provably valid probabilistic predictor, which has close connections to the powerful conformal prediction machinery, and we illustrate this construction in regression and classification applications.

Keywords and phrases: classification; conformal prediction; plausibility contour; random sets; regression

1 Introduction

Data-driven prediction of future observations is a fundamental problem. Here our focus is on applications where the data $Z = (X, Y)$ consists of explanatory variables $X \in \mathbb{X} \subseteq \mathbb{R}^d$, for some $d \geq 1$, and a response variable $Y \in \mathbb{Y}$. That is, we observe a collection $Z^n = \{Z_i = (X_i, Y_i) : i = 1, \ldots, n\}$ of n pairs from an exchangeable process. The two most common examples are *regression* and *classification*, where $\mathbb{Y}$ is an open subset and finite subset of $\mathbb{R}$, respectively. We consider both cases in what follows. The prediction problem corresponds to a case where we are given a value x_{n+1} of the next explanatory variable X_{n+1}, and the goal is to predict the corresponding future response $Y_{n+1} \in \mathbb{Y}$.

By "prediction" we mean quantifying uncertainty about Y_{n+1} in a data-dependent way, i.e., depending on the observed data Z^n and the given value x_{n+1} of X_{n+1}. One

[*] This is an extended version of the 2021 International Symposium on Imprecise Probability Theory and Applications (ISIPTA) proceedings paper, Cella and Martin (2021b).

[†] Department of Statistics, North Carolina State University; `lolivei@ncsu.edu`, `rgmarti3@ncsu.edu`

perspective on prediction uncertainty quantification is the construction of a suitable family of prediction sets representing collections of sufficiently plausible values for Y_{n+1}; see, e.g., Vovk et al. (2005), Campi et al. (2009), Kuleshov et al. (2018), and Equation (19) below. While prediction sets are practically useful, there are prediction-related tasks that they cannot perform, in particular, it cannot assign degrees of belief (or betting odds, etc.) to all relevant assertions or hypotheses "$Y_{n+1} \in A$," for $A \subseteq \mathbb{Y}$. An alternative approach is to develop what we refer to here as a *probabilistic predictor*, i.e., a probability-like structure (precise or imprecise probability) defined on $\mathbb{Y}$, depending on Z^n and x_{n+1}, designed to quantify uncertainty about Y_{n+1} by directly assigning degrees of belief to relevant assertions. The most common approach to probabilistic prediction is Bayesian, where a prior distribution for the model is specified and uncertainty is quantified by the posterior predictive distribution of Y_{n+1}, given Z^n and $X_{n+1} = x_{n+1}$. Other non-Bayesian approaches leading to predictive distributions include Lawless and Fredette (2005), Coolen (2006), Wang et al. (2012), and Vovk et al. (2018).

Before moving forward, it is important to distinguish between uncertainty quantification with prediction sets and with probabilistic predictors. One does not need a full (precise or imprecise) probability distribution to construct prediction sets and, moreover, sets derived from a probabilistic predictor are not guaranteed to satisfy the frequentist coverage probability property that warrants calling them genuine "prediction sets." Therefore, the motivation for going through the trouble of constructing probabilistic predictor, Bayesian or otherwise, must be that there are important prediction-related tasks that prediction sets cannot satisfactorily handle. In other words, the belief assignments provided by a (precise or imprecise) probability must be a high priority. Strangely, however, the reliability of probabilistic predictors is only ever assessed in terms of (asymptotic) coverage probability properties of their corresponding prediction sets. Our unique perspective is that, since belief assignments are a priority, there ought to be a way to directly assess the reliability of a probabilistic predictor's belief assignments.

For prediction problems where only the (response) variables $Y_1, \ldots, Y_n$ are observed, Cella and Martin (2022) introduced a notion of validity for probabilistic predictors. Roughly, their validity condition requires that the subsets $A \subseteq \mathbb{Y}$ to which the probabilistic predictor tends to assign large numerical degrees belief are the same as those that tend to contain Y_{n+1}. The point being that such a constraint ensures that the belief assignments made by the probabilistic predictor are not systematically misleading. Here we extend their notion of validity to the case where explanatory variables are present, and the precise definitions are given below in Definitions 1–2. It turns out these notions of validity have some important consequences, imposing certain constraints on the mathematical structure of the probabilistic predictor. Indeed, we argue in Section 3 (see, also, Corollary 1 in Section 4) that validity can only be achieved by probabilistic predictors that take the form of an imprecise probability distribution. Section 2 provides a preview of the formal definition of validity and offers empirical support for the claim that precise probabilistic predictors cannot be valid.

After formally introducing these notions of validity in Section 3, we explore their behavioral and statistical consequences. First, we show that even the weaker validity property in Definition 1 implies that the probabilistic predictor avoids (a property stronger than) the familiar sure loss property in the imprecise probability literature, hence is not internally irrational from a behavioral point of view. We go on to show that prediction-related procedures, e.g., tests and prediction regions, derived from (uniformly) valid probabilistic predictors control frequentist error probability. The take-away message is that a (uniformly) valid probabilistic predictor provides the "best of both worlds"—it simultaneously achieves desirable behavioral and statistical properties.

Given the desirable properties of a valid probabilistic predictor, the natural question is *how to construct one?* The probabilistic predictor we construct here is largely based on the general theory of valid *inferential models* (IMs) as described in Martin and Liu (2013, 2015b). Martin and Liu's construction usually assumes a parametric model but, here, we aim to avoid such strong assumptions. For this, we use a particular extension of the so-called *generalized IM* approach developed in Martin (2015, 2018). The basic idea is that a link/association between observable data, quantities of interest, and an unobservable auxiliary variable with known distribution can be made without fully specifying the data-generating process. In Section 5, we develop a valid IM construction that assumes only exchangeability of the observed data process, no parametric model assumptions required. There, in Theorem 1, we establish that this general IM-based probabilistic predictor construction achieves the (uniform) validity property. The specifics of this construction are presented in Section 6, in the context of regression. Section 7 considers the classification problem, and we show that the discreteness of Y in classification problems may cause the IM random set output, from which the probabilistic predictor is derived, to be empty with positive probability. Two possible adjustments are provided, with the one based on suitably "stretching" the random set being most efficient.

An important observation is that parallels can be drawn between our proposed IM construction and the conformal prediction approach put forward in Vovk et al. (2005) and elsewhere. This is interesting for at least two reasons.

- It demonstrates that one does not necessarily need "new methods" to construct probabilistic predictors to achieve the desired (uniform) validity property, just an appropriate re-interpretation of the output returned by certain existing methods. In particular, our proposed IM construction returns a possibility measure whose contour function is the transducer derived from an appropriate conformal prediction algorithm. Consequently, all we need is the corresponding conformal prediction algorithm to achieve our goals.
- However, there would be a variety of different ways the conformal prediction algorithm could be re-interpreted as a probabilistic predictor, e.g., as a precise probability distribution or one of several different imprecise probability distributions. Our developments here reveal that the appropriate re-interpretation, the one that leads to (uniform) validity, is by treating the conformal transducer as the contour function that defines a possibility measure.

These points, along with some other concluding remarks, are given in Section 8.

2 Prediction validity: a preview

To help clarify the difference between the traditional notions of uncertainty quantification in (probabilistic) prediction and the notions we have in mind here, we consider a relatively simple example for illustration, one in which there are no covariates. That is, suppose we have a sequence of real-valued observables $Y_1, Y_2, \ldots$ and, based on the observations $Y^n = y^n$, the goal is to predict Y_{n+1} in a probabilistic way. One standard way to approach this is to construct a Bayesian predictive distribution. This requires specification of a prior distribution over the space of models, an updating step whereby the prior distribution is updated to posterior distribution via Bayes's theorem in light of the observation $Y^n = y^n$, and then that the posterior is converted into a predictive distribution for Y_{n+1}, which we will denote by Π^n; keep in mind that Π^n is a function data Y^n.

Of course, there are a number of different summaries one can extract from the predictive distribution Π^n. Very often, the only summary considered is a prediction interval, e.g., the smallest set A such that $\Pi^n(A)$ is no less than $1-\alpha$, for some specified level $\alpha \in (0,1)$. Let this prediction interval be denoted by $C_\alpha(y^n)$. By construction, $C_\alpha(y^n)$ has posterior predictive probability at least $1-\alpha$, but one typically wants to give this a frequentist interpretation, to conclude that $C_\alpha(Y^n)$ has prediction coverage probability at least the nominal level $1-\alpha$; see Equation (2) below. In many cases, the Bayesian prediction interval will satisfy this frequentist coverage probability property, at least approximately. But one might ask: if the goal is to get a prediction interval that attains certain frequentist coverage properties, then why go to the trouble of constructing a full posterior predictive distribution for Y_{n+1}? There are certain advantages to quantifying uncertainty with a full predictive distribution, so these deserve exploration.

At a fundamental level, it is the predictive probabilities, i.e., $\Pi^n(A)$ for various A, that should be meaningful to the data analyst who opted to construct Π^n. That is, values of $\Pi^n(A)$ that are large (resp. small) should suggest that the data show strong (resp. weak) support for the claim "$Y_{n+1} \in A$." Therefore, based on the observed data and his predictive distribution construction, the data analyst would be inclined to conclude that the aforementioned claims will hold for A with large $\Pi^n(A)$ and will not hold for those with small $\Pi^n(A)$. But if the data analyst is thinking about the predictive distribution as a *method* for prediction, rather than summarizing his personal beliefs about Y_{n+1}, then he should care about the reliability of this method. This begs the following question: as a function of Y^{n+1}, do the two events $\{\Pi^n(A) \text{ is small}\}$ and $\{Y_{n+1} \notin A\}$ tend to happen simultaneously for all the relevant A's? If not, then the predictive distribution, treated as a method for predictive inference, lacks reliability in the sense that there is risk of erroneous predictions. Put differently, suppose the data analyst is a gambler who uses his $\Pi^n(A)$ values to set prices for \$1 bets on the uncertain outcomes "$Y_{n+1} \in A$." Then a lack of reliability in the sense described above implies existence of some A for which the gambler tends to assign a low price to "$Y_{n+1} \in A$" and have to pay out \$1. Of course, a tendency to lose \$1 on low-priced bets can easily lead to ruin. The details in the above discussion will all be formalized in Section 3.

Do the common Bayesian predictive distributions achieve this sort of reliability? One way to assess this would be to consider the function

$$f(\alpha) = \mathsf{P}\{\Pi^n(A) \leq \alpha,\, Y_{n+1} \in A\}, \quad \alpha \in (0,1), \tag{1}$$

where P denotes the joint distribution of Y^{n+1}. This function depends implicitly on A, on the chosen construction $y^n \mapsto \Pi^n$ of the predictive distribution, and on the underlying distribution P. It will be argued below that reliability or, rather, *validity* of a probabilistic predictor corresponds to

$$f(\alpha) \leq \alpha \quad \text{for all } (\alpha, n, A, \mathsf{P}).$$

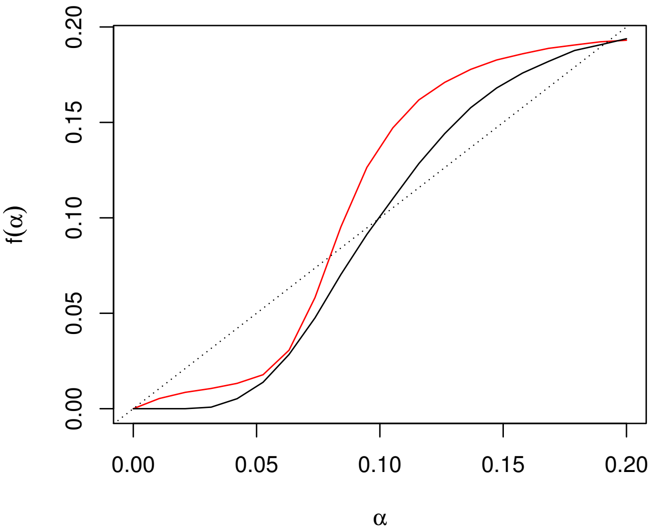


Figure 1: Plot of the function f in (1) corresponding to the two Bayesian predictive distributions (and other settings) as described in the text.

Figure 1 plots the function f for $n = 5$, $A = [3, 5]$, and P corresponding to iid $\mathsf{Unif}(-5, 5)$ random variables. This is done for two Bayesian predictive distributions:

- a parametric version based on a simple iid normal model with a conjugate normal–inverse gamma prior, leading to a suitable Student-t predictive distribution;
- and a nonparametric version based on the predictive distribution from a Dirichlet process mixture of normals model (e.g., Ghosh and Ramamoorthi 2003, Ch. 5); in fact, this is based on the (non-asymptotic) approximation in Hahn et al. (2018).

In both cases, we clearly see that there is an interval of α values at which the condition "$f(\alpha) \leq \alpha$" fails, hence the Bayesian predictive distribution is not valid in the sense described vaguely above, and more precisely in Section 3. That this is not specific to this example and these choices of predictive distribution is established in Corollary 1 below. That precise predictive distribution can fail to be reliable in the sense above motivates our investigation into other probabilistic predictor constructions that are reliable, which necessarily must take the form of imprecise probabilities.

Of course, Bayesian predictive distributions are not advertised to be reliable in this sense, so various excuses can be given, e.g., that the condition is too strong, that a different choice of prior distribution would perform better, etc. To us, however, the fact that a Bayesian probabilistic predictor (or any other construction based on a precise probability distribution for that matter) is sure to put the data analyst and gambler at risk of systematically misleading conclusions and ruin, respectively, is a serious concern. If precise probability were our only option, then of course we would just have to live with the aforementioned risk. Here we show, however, that suitably incorporating imprecision into the construction can nullify these risks without sacrificing on the other desirable properties that probabilistic prediction affords.

3 Prediction validity

3.1 Setup

The goal here is to formalize the ideas discussed in Section 2 above. Recall that the present paper is concerned with prediction in supervised learning problems, so we assume there is an exchangeable process $Z^\infty = (Z_1, Z_2, \ldots)$ with distribution P, where each Z_i is a pair $(X_i, Y_i) \in \mathbb{Z} = \mathbb{X} \times \mathbb{Y}$. As is customary, "$\mathsf{P}(Z^n \in B)$" is understood to mean the marginal probability for the event "$Z^n \in B$" derived from the joint distribution of Z^∞ under P. The distribution P is completely unknown, except that it belongs to the user-defined model $\mathscr{P}$ consisting of exchangeable distributions. As is typical in statistical learning applications, we want to avoid strong model assumptions, which amounts to assuming $\mathscr{P}$ is large, i.e., is not indexed by a finite-dimensional parameter. Given the observed data Z^n and a value x_{n+1} of X_{n+1}, the goal is to reliably predict the corresponding Y_{n+1}. As discussed in Sections 1–2, a common strategy is to construct a *prediction set* aimed to achieve the nominal frequentist coverage probability. That is, a collection of functions $C_{n,\alpha}$, from $\mathbb{Z}^n \times \mathbb{X}$ to subsets of $\mathbb{Y}$, indexed by $\alpha \in [0,1]$ and $n \geq 1$, defines a family of $100(1-\alpha)\%$ prediction sets for Y_{n+1} if

$$\inf_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{C_{n,\alpha}(Z^n, X_{n+1}) \ni Y_{n+1}\} \geq 1 - \alpha, \quad \text{for all } (\alpha, n). \tag{2}$$

However, as discussed in Section 2, a more "complete" uncertainty quantification about Y_{n+1} may be desired, beyond prediction sets. To formalize this, we follow Cella and Martin (2022) and define a *probabilistic predictor* as a map $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$, where $(\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is a pair of lower and upper predictive probabilities for the corresponding Y_{n+1}; for notational simplicity, the probabilistic predictor's dependence on the observed data z^n is encoded in the superscript "n" only. Then uncertainty quantification about Y_{n+1}, given z^n and $X_{n+1} = x$, is provided by the function $A \mapsto (\underline{\Pi}_x^n(A), \overline{\Pi}_x^n(A))$.

We are defining the probabilistic predictor for all n, but it could be that some minimum sample size is needed in order to properly define it. For example, if some standardization procedure is being employed, then it would be necessary to have n large enough to estimate standard errors. As a rule in what follows, if n is smaller than the necessary sample size, then we will silently take the probabilistic predictor to be vacuous, i.e., assign lower and upper probabilities 0 and 1, respectively, to every assertion.

What kind of mathematical form should the function $A \mapsto (\underline{\Pi}_x^n(A), \overline{\Pi}_x^n(A))$ take? Let $\mathcal{A}$ denote a σ-algebra of subsets of $\mathbb{Y}$ that are measurable with respect to the (common) marginal of the Y_i's under P. We will assume that $\mathcal{A}$ is rich enough to contain the singletons, e.g., like the Borel σ-algebra. Then the lower and upper probabilities are capacities defined on $\mathcal{A}$, i.e., monotone set functions, taking value 1 at $\mathbb{Y}$ and value 0 at $\varnothing$. However, being "lower" and "upper" suggests a link between the two. We formalize by requiring that, for each z^n and new value x of the feature X_{n+1}, the upper probability $\overline{\Pi}_x^n$ for Y_{n+1} is sub-additive; in particular, for any disjoint A and A', the upper probability satisfies $\overline{\Pi}_x^n(A \cup A') \leq \overline{\Pi}_x^n(A) + \overline{\Pi}_x^n(A')$. Then the lower probability $\underline{\Pi}_x^n$ is defined as the dual or conjugate to the upper probability,

$$\underline{\Pi}_x^n(A) = 1 - \overline{\Pi}_x^n(A^c), \quad A \in \mathcal{A}, \tag{3}$$

and from sub-additivity is follows that

$$\underline{\Pi}_x^n(A) \leq \overline{\Pi}_x^n(A), \quad A \in \mathcal{A},$$

hence the name "lower" and "upper" probabilities. Ordinary or precise probabilities are (sub)additive so they satisfy these condition with $\underline{\Pi}_x^n \equiv \overline{\Pi}_x^n$. Moreover, all of the standard imprecise probability models—belief functions, possibility measures, lower/upper previsions—satisfy these conditions, so our assumptions corresponding to no loss of generality. Since we will be interested in the statistical properties of the probabilistic predictor as functions of the data, we will assume that $(Z^n, X_{n+1}) \mapsto (\underline{\Pi}_{X_{n+1}}^n(A), \overline{\Pi}_{X_{n+1}}^n(A))$ is measurable for each $n \geq 1$ and for each $A \in \mathcal{A}$.

The interpretation of the probabilistic predictor's output is subjective and goes as follows. For given data z^n and a new value x of the feature X_{n+1}, the lower and upper probabilities represent

$$\begin{aligned}\underline{\Pi}_x^n(A) &= \text{maximum buying price for the gamble } \$1(Y_{n+1} \in A) \\ \overline{\Pi}_x^n(A) &= \text{minimum selling price for the gamble } \$1(Y_{n+1} \in A),\end{aligned}$$

where $1(B)$ denotes the indicator of the event B. Therefore, based on data z^n and new feature x, if the investigator's $\underline{\Pi}_x^n(A)$ is large, then she would be inclined to buy the gamble $\$1(Y_{n+1} \in A)$, whereas, if her $\overline{\Pi}_x^n(A)$ is small, then she would be inclined to sell the gamble $\$1(Y_{n+1} \in A)$; otherwise, she might choose to neither buy nor sell the gamble. For this reason, $\underline{\Pi}_x^n(A)$ measures the subjective degree of belief and $\overline{\Pi}_x^n(A)$ the plausibility of the event "$Y_{n+1} \in A$." Below we introduce an element of objectivity through a requirement that its predictions be reliable in a statistical sense.

3.2 Definition

So far, we have imposed minimal mathematical constraints on the probabilistic predictor, plus its interpretation is subjective, so virtually no construction can be ruled out at this point. However, the probabilistic predictor's practical utility requires that the uncertainty quantification derived from it be reliable in a certain sense. The particular sense we have in mind is *statistical.* That is, we require that inferences drawn based on the probabilistic predictor not be systematically misleading. Based on the interpretations of the lower and upper probabilities described above, events of the general form

$$\{(z^n, x_{n+1}, y_{n+1}) : \underline{\Pi}_{x_{n+1}}^n(A) \text{ is large and } y_{n+1} \not\in A\}$$

and

$$\{(z^n, x_{n+1}, y_{n+1}) : \overline{\Pi}_{x_{n+1}}^n(A) \text{ is small and } y_{n+1} \in A\},$$

should they occur, put the investigator at risk of making erroneous predictions and incurring losses, monetary or otherwise. To protect the investigator from this risk, we impose the following condition on probabilistic predictors, ensuring that the aforementioned undesirable, risk-creating events are controllably rare.

Definition 1. The probabilistic predictor $(Z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is *valid* if one and, hence, both of the following equivalent conditions hold:

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha\,,\, Y_{n+1} \notin A\} \leq \alpha, \quad \text{for all } (\alpha, n, A) \tag{4}$$

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha\,,\, Y_{n+1} \in A\} \leq \alpha, \quad \text{for all } (\alpha, n, A). \tag{5}$$

Here "for all (α, n, A)" is short for "for all $\alpha \in [0,1]$, all $n \geq 1$ and all $A \in \mathcal{A}$." The two conditions are equivalent by the duality in (3) and the "for all A" clause.

The key point, again, is that validity ensures the probabilistic predictor will not tend to assign small upper probability to assertions about Y_{n+1} that happen to be true, or large lower probability to assertions about Y_{n+1} that happen to be false. Practically, this ensures that the data analyst is not making systematically misleading predictions. Such assurances are also fundamentally important to the logic of statistical reasoning. Following Fisher (1973, p. 42), what makes an observation leading to, say, a small value of $\overline{\Pi}_{x_{n+1}}^n(A)$ informative about the claim "$Y_{n+1} \in A$" is that a logical disjunction is created: *either the claim does not hold or a small-probability event has occurred.* Since small-probability events rarely occur, if we observe a small value of $\overline{\Pi}_{x_{n+1}}^n(A)$, then we are inclined to conclude that $Y_{n+1} \notin A$. Validity also has a number of interesting and practically relevant consequences, which we explore in Section 4.

Before moving on, we should mention some connections with certain notions of "frequency calibration" in the imprecise probability literature. In particular, using our terminology and notation, Denœux (2006) defines a probabilistic predictor to have a "$100(1-\alpha)\%$ confidence property," for a fixed $\alpha \in [0,1]$, if

$$\mathsf{P}\bigl\{\overline{\Pi}_{X_{n+1}}^n(A) \geq \mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1}) \text{ for all } A\bigr\} \geq 1 - \alpha.$$

This and other variations are discussed more recently in Denœux and Li (2018). Obviously, since the event on the left-hand side does not explicitly depend on α, it must be that the probabilistic predictor depends implicitly on the specified α value, and various approaches to incorporate this α-dependence so that the above property can be achieved are given in the aforementioned references. The key observation is that calibration requires some relation between the probabilistic predictor for Y_{n+1} and the true conditional distribution of Y_{n+1}. In particular, the prediction upper probability ought to dominate the true conditional probability in some sense. A similar dominance appears in our definition of validity, but a slight reformulation is needed. Using iterated expectation, by conditioning on (Z^n, X_{n+1}), it is easy to see that (5) is equivalent to

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{E}\bigl[1\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha\}\, \mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})\bigr] \leq \alpha, \tag{6}$$

where the expectation is with respect the same P over which the supremum is taken. That is, our notion of validity implies that, when restricted to data sets (Z^n, X_{n+1}) for which $\overline{\Pi}_{X_{n+1}}^n(A)$ is small, the true conditional probability $\mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})$ cannot be any bigger on average. Incidentally, there are other notions of calibration/validity in the literature that concern matching up posited predictive distributions with the true probabilities in an average sense, not so unlike what (6) achieves. See, for example, the calibration property of Venn–Abers predictors in Vovk and Petej (2014) and the calibration safety property in Grünwald (2018).

3.3 A stronger notion

That the calibration property imposed in (5) is required to hold for all $A \subseteq \mathbb{Y}$ might seem overly strong, but it turns out that there is an even stronger property that is particularly useful and can be readily attained. To state this new property, however, we need some additional notation. Define the probabilistic predictor's *plausibility contour* as

$$\pi_x^n(y) = \overline{\Pi}_x^n(\{y\}), \quad x \in \mathbb{X}, \quad y \in \mathbb{Y}. \tag{7}$$

This is just the upper probability—or plausibility—assigned to singleton assertions about Y_{n+1} of the form $A = \{y\}$, for generic $y \in \mathbb{Y}$. In general, the plausibility contour is just one of the probabilistic predictor's many features. But in the important special case where the probabilistic predictor has the mathematical property of *consonance*, the plausibility contour actually determines the entire probabilistic predictor. We will discuss this latter point further below.

Definition 2. The probabilistic predictor $(Z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is *uniformly valid* if

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\pi_{X_{n+1}}^n(Y_{n+1}) \leq \alpha\} \leq \alpha, \quad \text{for all } (\alpha, n), \tag{8}$$

The condition (8) is familiar, at least when connections are drawn to other contexts. In particular, (8) closely resembles the properties satisfied by p-values from hypothesis testing in classical statistics. It is also effectively the same as the so-called *fundamental frequentist principle*, or *FFP*, in Walley (2002). But there are still some unanswered questions: in what sense is this definition stronger than that in Definition 1, and why do we call this "uniform" validity? The following lemma helps us to answer both.

Lemma. *Uniform validity in the sense of Definition 2 is equivalent to the probabilistic predictor satisfying the following two properties:*

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha \textit{ and } Y_{n+1} \notin A \textit{ for some } A\} \leq \alpha, \quad \textit{for all } (\alpha, n) \tag{9}$$

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha \textit{ and } Y_{n+1} \in A \textit{ for some } A\} \leq \alpha, \quad \textit{for all } (\alpha, n). \tag{10}$$

Proof. That both probabilities on the left-hand sides of (9) and (10) are equal to the left-hand side of (8) follows from the probabilistic predictor's monotonicity property. □

That uniform validity in the sense of Definition 2 is stronger than validity in the sense of Definition 1 can now be readily seen. Indeed, the "for some A" inside the probability statement in (10) is effectively a union of A-dependent events like those in (5) over all A. So if the union over A of these A-dependent events has probability bounded by α, then so would any individual event in that union. This also explains our choice to describe this as "uniform validity." That is, instead of a "$\leq \alpha$" bound that holds for each individual assertion A, it now must hold simultaneously or uniformly over all such A.

This generalization is important for several reasons. One of those reasons is technical; see Proposition 2 below. Another concerns the point that when the "for all A" clause on the outside of the probability statement in (5) is moved to the inside, the choice of A can be data-dependent. To see why this data-dependence might be relevant, consider a gambling scenario in which the agent's opponents have access to the data (z^n, x) at the time of prediction. This allows the opponent to use the data to make strategic choices about which assertions A to negotiate with the agent. Of course, if the agent's opponents can make these more sophisticated data-dependent plays while he is only able to control errors for assertions specified in advance, then that puts him at risk. Uniform validity, however, protects the agent from this more subtle type of risk.

Although we currently lack a formal proof, our experience suggests that only *consonant* (Shafer 1976, Ch. 10) probabilistic predictors can achieve uniform validity. The reason being that, if the plausibility contour, π_x^n, in (7), is restricted in the sense that it cannot attain values arbitrarily close to 1, then the stochastically-no-smaller-than-uniform condition in (8) likely cannot hold. And it is precisely this arbitrarily-close-to-1 property that determines consonance; that is, a probabilistic predictor is consonant if and only if its plausibility contour satisfies

$$\sup_y \pi_x^n(y) = 1, \quad \text{for all } (z^n, x). \tag{11}$$

In this case, the probabilistic predictor takes the mathematical form of a possibility measure (Dubois and Prade 1988), and is determined by its contour function through the relationship

$$\overline{\Pi}_x^n(A) = \sup_{y \in A} \pi_x^n(y), \quad A \in \mathcal{A}. \tag{12}$$

Since the lower and upper probabilities being determined by a single point-function, as opposed to genuine set-functions, consonance amounts to a substantial simplification of the probabilistic predictor. For our purposes here, and for statistical inference in general (Martin 2021), this simplification comes with no loss of generality or flexibility.

4 Implications of prediction validity

4.1 Behavioral

Despite our focus on frequentist-style properties, validity has some important behavioral consequences, à la de Finetti, Walley, and others. Towards this, define

$$\underline{\gamma}_n(A) = \inf_{(z^n,x) \in \mathbb{Z}^n \times \mathbb{X}} \underline{\Pi}_x^n(A) \quad \text{and} \quad \overline{\gamma}_n(A) = \sup_{(z^n,x) \in \mathbb{Z}^n \times \mathbb{X}} \overline{\Pi}_x^n(A),$$

the lower/upper probabilistic predictor evaluated at A, optimized over all of its data inputs; recall that $\underline{\Pi}_x^n$ and $\overline{\Pi}_x^n$ depend implicitly on an argument z^n. An especially poor specification of prediction probabilities is a situation in which, for some $A \subseteq \mathbb{Y}$,

$$\underline{\gamma}_n(A) > \inf_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A) \quad \text{or} \quad \overline{\gamma}_n(A) < \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A). \tag{13}$$

We will refer to this as (one-sided) *contraction*. Ideally, the probabilistic predictor would mimic the true conditional probability at least in the sense that its average over data inputs would not be far from the true marginal probability. So a situation like in (13), where the probabilistic predictor might be uniformly bounded away from the true marginal probability, is a sign of potential trouble. For example, if your $\overline{\Pi}_x^n(A)$ is smaller than the upper bound on the marginal probability of A, uniformly in (z^n, x), i.e., *no matter what data is observed*, then arguably you should have had a tighter bound on your marginal probability to start. Inconsistencies like this factor in to the behavioral properties of the probabilistic predictor, and (imprecise) probabilities more generally. For example, the *sure loss* property—see Condition (C7) in Walley (1991, Sec. 6.5.2) or Definition 3.3 in Gong and Meng (2021)—corresponds to an extreme version of contraction where

$$\underline{\gamma}_n(A) > \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A) \quad \text{or} \quad \overline{\gamma}_n(A) < \inf_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A).$$

In a gambling context, an inconsistency as severe as in the above display can be leveraged by your opponent to make you a sure loser. We show below, in Proposition 1, that validity and one-sided contraction are incompatible; in particular, validity implies no sure loss.

Although (13) is still a rather strong condition, corresponding to a poor prediction probability specification, there are practically relevant cases where (13) holds and creates a genuine risk. We discuss this below following the proof.

Proposition 1. *Suppose that the probabilistic predictor, $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$, suffers from one-sided contraction in the sense that* (13) *holds for some $A \subseteq \mathbb{Y}$. Then validity in the sense of Definition 1 fails.*

Proof. We present the argument here for the case where $\overline{\gamma}_n(A) < \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A)$; the argument for the $\underline{\gamma}_n(A)$ bound is very similar. For the assertion A in (13), define

$$\xi_n(A, \alpha) = \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha \, , \, Y_{n+1} \in A\},$$

so that (5) is equivalent to

$$\xi_n(A, \alpha) \leq \alpha \quad \text{for all } (A, \alpha, n). \tag{14}$$

By (6), we have

$$\xi_n(A, \alpha) = \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{E}\big[1\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha\} \, \mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})\big].$$

Since $\overline{\Pi}_{X_{n+1}}^n(A) \leq \overline{\gamma}_n(A)$ by definition, we get

$$1\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha\} \geq 1\{\overline{\gamma}_n(A) \leq \alpha\}.$$

From the alternative representation of $\xi_n(A, \alpha)$, since the lower bound in the above display is constant, it follows that

$$\xi_n(A, \alpha) \geq 1\{\overline{\gamma}_n(A) \leq \alpha\} \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A). \tag{15}$$

According to (13), there exists an $A \subseteq \mathbb{Y}$ and a threshold $\alpha \in [0, 1]$ such that

$$\overline{\gamma}_n(A) < \alpha < \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A).$$

Then from (15), with this choice of (A, α),

$$\xi_n(A, \alpha) \geq \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A) > \alpha.$$

Then (14) and, hence, (5) fails, so the claim follows. □

Our one-sided contraction (13) also resembles the (C8) portion of the *coherence* property in Walley (1991, Sec. 6.5.2), but it is missing the (C9) portion. Therefore, it appears that validity is not enough to imply coherence as advocated for by Walley and others.

As discussed above, a very common strategy is one in which the probabilistic predictor is an ordinary probability, i.e., $(z^n, x) \mapsto \Pi_x^n$, where Π_x^n is a (precise) probability distribution on $\mathbb{Y}$. The illustration presented in Section 2 suggests that validity might fail when the probabilistic predictor is a (precise) probability distribution. Of course, if the true distribution is known, i.e., if $\mathscr{P}$ is a singleton, then setting Π_n^x equal to the true conditional distribution would be valid; see (6). However, when $\mathscr{P}$ is big, as assumed here, we cannot expect that a probability distribution can accommodate both the inherent variability in the response and the uncertainty about the underlying distribution. So we should anticipate that imprecision is needed in order for predictions to be valid in the sense of Definition 1. The discussion below formalizes this claim.

A precise probabilistic predictor cannot accommodate uncertainty about the model by assigning a range of probabilities for a given assertion. As such, the probabilities $\Pi_n^x(A)$ will tend to be strictly between 0 and 1. But if the model $\mathscr{P}$ is large, we fully expect the infimum and supremum of $\mathsf{P}(Y_{n+1} \in A)$ over $\mathscr{P}$ to be 0 and 1, respectively. Therefore, if there exists an assertion A such that

$$\inf_{(z^n,x) \in \mathbb{Z}^n \times x} \Pi_x^n(A) > 0 \quad \text{or} \quad \sup_{(z^n,x) \in \mathbb{Z}^n \times x} \Pi_x^n(A) < 1, \tag{16}$$

with inequalities strict, then (13) holds and, by Proposition 1, validity fails. One situation in which the inequalities are strict for some A is when

$$\{\Pi_x^n : (z^n, x) \in \mathbb{Z}^n \times \mathbb{X}\} \quad \text{is a tight collection of distributions.} \tag{17}$$

Roughly speaking, tightness prevents the collection of distributions from "drifting off to infinity," keep at least some amount of probability mass to the interior of $\mathbb{Y}$. Tightness always holds for compact $\mathbb{Y}$, at least in all practical cases; for non-compact $\mathbb{Y}$, it would need to be verified case-by-case, using specific features of the map $(z^n, x) \mapsto \Pi_x^n$. In any case, tightness leads to strict contraction (16) for some A, which implies one-sided contraction, which implies validity fails.

Corollary 1. *If the probabilistic predictor, $(z^n, x) \mapsto \Pi_x^n$, is a precise probability distribution that satisfies* (17)*, then it is not valid.*

Proof. A direct consequence of Proposition 1. □

The above result establishes a version of the *false confidence theorem* (Balch et al. 2019; Martin 2019) in the context of prediction. It says roughly the following: only probabilistic predictors that take the form of an imprecise probability can be valid in the sense of Definition 1. We do not expect the tightness condition (17) is essential to the conclusion of Corollary 1, but we currently are not aware of a direct proof.

4.2 Statistical

Here we consider some more classical frequentist-style prediction tasks. First, consider testing certain "hypotheses" about Y_{n+1}. For example, an investor may want to sell a certain asset when its price exceeds some fixed level, say $y^\star$. So he would like to assess the plausibility of an assertion or hypothesis of the form "$Y_{n+1} \in A$," for $A = [0, y^\star]$ and, in particular, decide if the new price being below the $y^\star$ threshold is too plausible to warrant taking quick action to sell. We show below that the test

$$\text{reject "}Y_{n+1} \in A\text{" if and only if } \overline{\Pi}_x^n(A) \leq \alpha, \tag{18}$$

derived from a valid probabilistic predictor controls the error probability at level α.

A more common prediction-related task is the construction of a prediction set, i.e., a set of sufficiently plausible values for Y_{n+1} given the observed data. A natural way to construct a prediction set from a probabilistic predictor is

$$C_{n,\alpha}(z^n, x) = \{y : \pi_x^n(y) > \alpha\}, \tag{19}$$

where π_x^n is the plausibility contour (7) based on (z^n, x). Compare this to a Bayesian highest predictive density region. The following proposition shows that uniform validity implies that this is a genuine $100(1-\alpha)\%$ prediction set in the sense that its frequentist coverage probability is at least the advertise/nominal level $1 - \alpha$.

Proposition 2. (a) *If the probabilistic predictor is valid in the sense of Definition 1, then the test described in* (18) *controls error rates at level α in the sense that*

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{\textit{test based on } (Z^n, X_{n+1}) \textit{ rejects and } Y_{n+1} \in A\} \leq \alpha.$$

(b) *If the probabilistic predictor is uniformly valid in the sense of Definition 2, then* (19) *defines a genuine $100(1-\alpha)\%$ prediction set in the sense that*

$$\sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}\{C_{n,\alpha}(Z^n, X_{n+1}) \not\ni Y_{n+1}\} \leq \alpha, \quad \textit{for all } (\alpha, n).$$

Proof. Part (a) is an immediate consequence of the definition of validity, in particular, (5). Part (b) follows directly from (8). □

Recall our conjecture that uniform validity is satisfied only for probabilistic predictors that are consonant, i.e., fully determined by their plausibility contour via (12). For consonant probabilistic predictors, the level sets of the plausibility contour, which are nested by definition, play an important role. In particular, this underlying nested structure allows us to re-express the prediction set in (19) in terms of the lower probability:

$$C_{n,\alpha}(z^n, x) = \bigcap\{A : \underline{\Pi}_x^n(A) \geq 1 - \alpha\}.$$

That is, $C_{n,\alpha}(z^n, x)$ can also be interpreted as the smallest assertion A about Y_{n+1} to which the probabilistic predictor assigns lower probability at least $1 - \alpha$.

5 Inferential models

A relevant question is how to construct a probabilistic predictor that achieves the (uniform) validity condition. One strategy would be through a *generalized Bayes* approach as advocated for in, e.g., Walley (1991, Sec. 6.4). That is, if $\mathscr{P}$ is the set of candidate joint distributions for the observables, the generalized Bayes rule would define an upper prediction probability as

$$\overline{\Pi}_x^n(A) = \sup_{\mathsf{P} \in \mathscr{P}} \mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1} = x), \quad A \in \mathcal{A}, \tag{20}$$

and corresponding lower probability by replacing sup by inf. That this satisfies validity in the sense of Definition 1 follows from the alternative formulation in (6). While this solution might have some appeal, there are also some reasons to be concerned. First, this would tend to be quite conservative, i.e., a large model space $\mathscr{P}$ implies a wide gap between lower and upper probabilities. Second, since generalized Bayes would not be consonant, it is doubtful that the desirable uniform validity in Definition 2 holds. So it is worth considering alternative constructions that might be more efficient.

An inferential model (IM) is a data-dependent probabilistic structure designed to quantify uncertainty about unknowns, like the probabilistic predictor described above. The difference is that, as the name suggests, IMs have traditionally focused on the *statistical inference* problem, where the unknowns are fixed quantities. IMs have connections to various other approaches to statistical inference, some that quantify uncertainties with ordinary probabilities, e.g., Bayesian inference, fiducial inference (Fisher 1935), and generalized fiducial inference (Hannig et al. 2016), and others with imprecise probabilities, e.g., Dempster–Shafer theory (Dempster 1967, 1968, 2008, 2014; Shafer 1976) and other belief function frameworks (Denœux 2014; Denœux and Li 2018). While there are some technical differences resulting from the unknown being fixed in the inference case and random in the prediction case, the common goal of providing valid uncertainty quantification is more or less the same. Therefore, we expect that the key ideas behind the construction of a valid IM for inference ought to be applicable to the prediction problem as well, modulo a few adjustments. Below we describe a construction of a probabilistic predictor that is valid in the sense described in Section 3.

The general IM construction is composed of three steps. The A-step *associates* the observable data and unknown quantity of interest with an unobservable auxiliary variable whose distribution is fully known. In the early work on IMs, this association was usually a complete description of the data-generating process. For example, suppose we have, say, n independent and identically distributed (iid) observations $Z_1, \ldots, Z_n$, collected into the vector Z^n, from a statistical model with unknown parameter θ. Then an association would effectively be a description of how to generate data Z^n from that model, i.e.,

$$Z^n = a(\theta, U^n),$$

where U^n would typically be a vector of iid latent/auxiliary variables with a known distribution, e.g., $\mathsf{Unif}(0,1)$. While such an association can always be written down, there are a few obstacles one might face when trying to complete the IM construction:

- When the dimension of U^n is greater than that of θ, as is typical, a dimension reduction step is recommended (Martin and Liu 2015a), but this can be nontrivial.

- The association itself requires (more than) a fully specified statistical model for data, which may not be available in the application at hand.

However, Martin (2015, 2018) showed that the A-step's requirements can be relaxed. All that is needed is an association that relates a function of both the data and the unknowns to an unobservable auxiliary variable. This idea has proved to be useful in a variety of classical (Cahoon and Martin 2020, 2021) and modern (Cella and Martin 2021a) inference problems, and here we develop a version suitable for prediction.

Once a generalized association has been set, the remaining steps of the (generalized) IM construction proceed exactly as described in, say, Martin and Liu (2013). Roughly, the P-step introduces a random set that aims to *predict* or guess the unobserved value of the auxiliary variable. Easy to arrange properties of this user-specified random set ensure that the guessing of the auxiliary variable is done in a reliable way, which turns out to be fundamental for validity. Next, the C-step *combines* the results of the A- and P-steps, yielding a new, data-dependent random set on the space where the quantity of interest resides. Finally, this random set's distribution determines lower and upper probabilities that can be used to assign degrees of belief and plausibility to any relevant assertion about the unknown quantities of interest. Below we describe the generalize IM construction in more detail for the prediction problem at hand.

For prediction, the unknown is Y_{n+1}, not a model parameter as in the formulations described above. So the kind of association needed is one that identifies a function of (Z^n, Z_{n+1}) that has a known distribution. Once found, the three-step (generalized) IM construction proceeds as follows.

A–step. Suppose there exists a function $\phi_n : \mathbb{Z}^n \times \mathbb{Z} \to \mathbb{R}$ such that the distribution, say, Q_n, of the random variable $\phi_n(Z^n, Z_{n+1})$ is known, i.e., does not depend on the unknown P. Then associate the observable data Z^n and the yet-to-be-observed Z_{n+1} with the unobservable auxiliary variable U as follows:

$$\phi_n(Z^n, Z_{n+1}) = U, \quad U \sim \mathsf{Q}_n. \tag{21}$$

For our case where $Z_{n+1} = (X_{n+1}, Y_{n+1})$ and interest is in Y_{n+1} for a given $X_{n+1} = x$, the association defines a set-valued mapping

$$(Z^n, x, u) \mapsto \mathbb{Y}_x^n(u) := \{y \in \mathbb{Y} : \phi_n(Z^n, (x, y)) = u\}.$$

P–step. Define a nested random set $\mathcal{U}$ (see below) on the space $\mathbb{U}$ of the auxiliary variable U, designed to reliably contain realizations of $U \sim \mathsf{Q}_n$ in the sense of (24) below. The distribution of the random set $\mathcal{U}$ will be denoted by R_n.

C–step. Combine the results of the A- and P-steps to get the data-dependent random set

$$\mathbb{Y}_x^n(\mathcal{U}) = \bigcup_{u \in \mathcal{U}} \mathbb{Y}_x^n(u) = \{y \in \mathbb{Y} : \phi_n(Z^n, (x, y)) \in \mathcal{U}\}.$$

Then the distribution of this new random set, derived from the distribution of $\mathcal{U}$, determines the probabilistic predictor for Y_{n+1}, i.e.,

$$\begin{aligned} \underline{\Pi}_x^n(A) &= \mathsf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \subseteq A\} \\ \overline{\Pi}_x^n(A) &= \mathsf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \cap A \neq \varnothing\}. \end{aligned} \tag{22}$$

Remark 1. If $\mathbb{Y}_x^n(\mathcal{U})$ is empty with positive R_n-probability, then some adjustment to the probabilistic predictor in (22) is needed. This will be relevant for the classification problem in Section 7.

The above construction is abstract for the purpose of generality. The challenge is in identifying the function ϕ_n, and examples will be given in Sections 6–7 below. Other examples were explored previously in Martin and Lingham (2016) where P was assumed to belong to a parametric family. Here, however, P is not indexed by a finite-dimensional parameter, so different techniques are required. The remainder of this section investigates the properties of the abstract probabilistic predictor construction above.

The random set $\mathcal{U}$ is assumed to be nested in the sense that, for any two sets in its support, one is a subset of the other. As a consequence, the derived probabilistic predictor is consonant; that is, its contour function, which is given by

$$\pi_x^n(y) = \mathsf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \ni y\}, \quad y \in \mathbb{Y}, \tag{23}$$

satisfies (11) and, hence, determines the entire probabilistic predictor through the relationship (12), i.e., $\overline{\Pi}_x^n(A) = \sup_{y \in A} \pi_x^n(y)$. As discussed in Section 3.3, consonance is important—perhaps necessary—for uniform validity.

It remains to establish that the probabilistic predictor resulting from the above construction is (uniformly) valid. This requires stating the conditions on $\mathcal{U}$ more precisely. Since the cases in the following sections involve an auxiliary variable U that is discrete, we will focus on the discrete case. First, define the random set's contour function

$$f(u) = \mathsf{R}_n(\mathcal{U} \ni u), \quad u \in \mathbb{U}.$$

Then the required link between Q_n and R_n is that

$$\text{if } U \sim \mathsf{Q}_n, \text{ then } f(U) \sim \mathsf{Unif}((n+1)^{-1}\mathscr{I}_{n+1}), \tag{24}$$

where $\mathscr{I}_{n+1} = \{1, \ldots, n, n+1\}$, so that this uniform distribution is discrete. With this link between the auxiliary variable's distribution Q_n and the random set's distribution R_n, we are ready to state and prove the main result.

Theorem 1. *If the random set $\mathcal{U}$ satisfies* (24)*, and if $\mathbb{Y}_{X_{n+1}}^n(\mathcal{U})$ is non-empty with R_n-probability 1 for P-almost all (Z^n, X_{n+1}), then the probabilistic predictor defined in* (22)*, or equivalently* (12)*, is uniformly valid in the sense of Definition 2.*

Proof. First, for Z^n and $Z_{n+1} = (X_{n+1}, Y_{n+1})$, set $U = \phi_n(Z^n, Z_{n+1})$. Then

$$\mathbb{Y}_{X_{n+1}}^n(\mathcal{U}) \ni Y_{n+1} \iff \mathcal{U} \ni U.$$

The R_n-probability of the left- and right-hand side events are $\pi_{X_{n+1}}^n(Y_{n+1})$ and $f(U)$, respectively, so these two random variables—the first as a function of $(Z^n, Z_{n+1}) \sim \mathsf{P}$ and the second as a function of $U \sim \mathsf{Q}_n$—have the same distribution. Equation (24) states that $f(U)$ is uniform and, therefore, so is $\pi_{X_{n+1}}^n(Y_{n+1})$. □

The non-emptiness condition is not necessary for validity, but some adjustment is needed to the definition in (22), as mentioned in Remark 1, to address this. We will discuss this below in the specific application to classification in Section 7. The requirement in (24) that $f(U) \sim \mathsf{Unif}((n+1)^{-1}\mathscr{I}_{n+1})$ can be relaxed in a certain sense without compromising validity. That is, validity also holds for any random set such that $f(U)$ is stochastically no smaller than $\mathsf{Unif}((n+1)^{-1}\mathscr{I}_{n+1})$. However, Cella and Martin (2022) show that the choice of $\mathcal{U}$ whose corresponding $f(U)$ is exactly uniformly distributed is most efficient. Fortunately, this is easy to arrange; see (28) below.

The following is an immediate consequence of the uniform validity conclusion above and the general results in Propositions 1–2 in the previous section.

Corollary 2. *Under the conditions of Theorem 1, the probabilistic predictor defined in* (22) *avoids sure loss in the sense of* (13) *and admits a prediction set $C_{n,\alpha}$ as in* (19) *that achieves the nominal frequentist prediction coverage probability.*

Consequently, the proposed probabilistic predictor construction achieves the desired subjective/behaviorist and objective/frequentist properties simultaneously. Two specific and practically relevant applications of this construction in the context of regression and classification will be presented in Section 6 and 7, respectively.

It is important to point out that the kind of validity being considered here is *marginal*, which is easiest to understand in the context of calibrated prediction sets as in (2). That is, the conditional coverage probability of the prediction set is

$$x_{n+1} \mapsto \mathsf{P}\{C_\alpha^n(x_{n+1}) \ni Y_{n+1} \mid X_{n+1} = x_{n+1}\},$$

a function of x_{n+1}. Then the validity property implies that the expected value of this function, with respect to the marginal distribution of X_{n+1} under P, is at least $1-\alpha$. This marginal coverage guarantee, of course, says nothing about the conditional coverage at any particular x_{n+1} values. Conditional validity is both challenging and practically relevant, and we discuss this briefly in Section 8.

6 Probabilistic prediction in regression

Recall that the A-step requires the specification of a real-valued function ϕ_n, such that the distribution of $\phi_n(Z^n, Z_{n+1})$ is known. Towards this, given $Z^{n+1} = (Z^n, Z_{n+1})$ consisting of the observable (Z^n, X_{n+1}) and the yet-to-be-observed Y_{n+1}, consider first a transformation $Z^{n+1} \to T^{n+1}$, defined by

$$T_i = \Psi(Z_{-i}^{n+1}, Z_i), \quad i \in \mathscr{I}_{n+1}, \tag{25}$$

where $Z_{-i}^{n+1} = Z^{n+1} \setminus \{(Y_i, X_i)\}$, and Ψ is a suitable real-valued function that compares Y_i to a prediction derived from Z_{-i}^{n+1} at X_i, being small if they agree and large if they disagree. For example, to each Z_{-i}^{n+1}, one could fit a regression model to get an estimated mean response $\hat\mu_{-i}^{n+1}(X_i)$ and take T_i as the corresponding absolute residual

$$T_i = \bigl|Y_i - \hat\mu_{-i}^{n+1}(X_i)\bigr|, \quad i \in \mathscr{I}_{n+1}. \tag{26}$$

The critical property of Ψ is that it be symmetric in the elements of its first vector argument. This symmetry guarantees that the assumed exchangeability in $Z_1, Z_2, \ldots$ is preserved when Z^{n+1} get mapped to T^{n+1}. As T_i depends on the entire data Z^{n+1}, we will write $T_i(Z^{n+1})$ where necessary to highlight that dependence. In regression, where the Y_i's are continuous and Ψ is non-constant on sets of Y^{n+1} with positive P-probability, like the one in (26), so that there are no ties, a well-known consequence of exchangeability of

$T_1, \ldots, T_{n+1}$ is that their ranks are marginally distributed according to $\mathsf{Unif}(\mathscr{I}_{n+1})$, the discrete uniform law on $\mathscr{I}_{n+1}$.

Having identified a function of (Z^n, Z_{n+1}) whose distribution is known, we can complete the A-step of the IM construction by writing a version of (21) as follows:

$$r(T_{n+1}) = U, \quad U \sim \mathsf{Unif}(\mathscr{I}_{n+1}), \tag{27}$$

where $r(\cdot)$ is the ascending ranking operator. The choice of T_{n+1} instead of any of the other T_i's in (27) is simply because T_{n+1} is the one that holds the to-be-predicted value, Y_{n+1}, in special status. Note that, while it appears this expression only depends on T_{n+1}, it does implicitly depend on all the T_i's and, hence, all of Z^{n+1}, through the ranking procedure. In summary, to complete the A-step, the only task for the data analyst is the specification of Ψ. It is worth to mention that, while validity of the probabilistic predictor is guaranteed for any suitable Ψ, choices of Ψ that fail to capture the structure of the problem at hand can lead to inefficiency.

For the P-step, the specification of a nested random set targeting the unobserved realization of the auxiliary variable U, introduced above, is needed. Consider

$$\mathcal{U} = \{1, 2, \ldots, U'\}, \quad U' \sim \mathsf{Unif}(\mathscr{I}_{n+1}). \tag{28}$$

It is straightforward to show that this random set satisfies the critical calibration property (24). Moreover, this choice also makes intuitive sense, as $\mathcal{U}$ always includes the value 1. This is desirable given the ascending ranking operator in (27) because it implies values of Y_{n+1} that make the residual T_{n+1} small will be assigned high plausibility.

Finally, in the C-step, $\mathcal{U}$ is combined with the u-indexed collection of sets

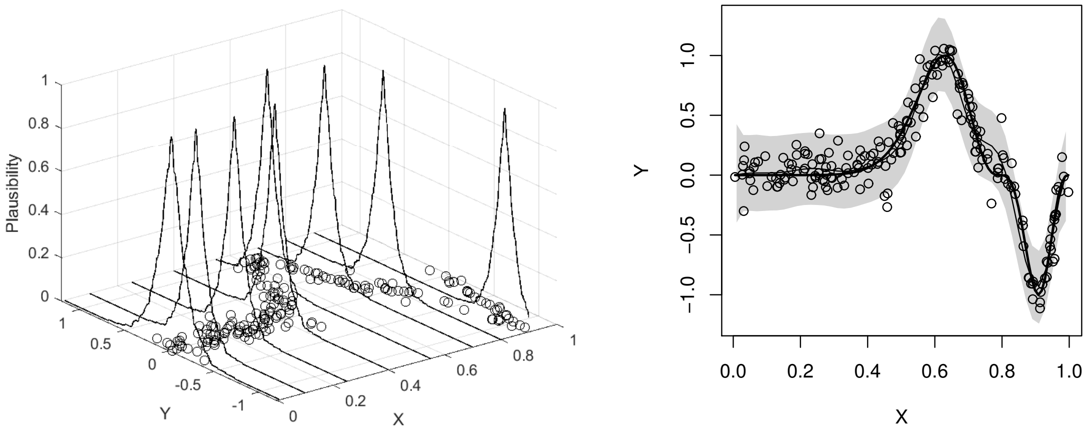
$$\mathbb{Y}^n_{x_{n+1}}(u) = \big\{y_{n+1} : r\big(T_{n+1}(z^{n+1})\big) = u\big\}$$

that arise from the association (27). Here and below, note that z^{n+1} consists of the observed z^n values with $z_{n+1} = (x_{n+1}, y_{n+1})$ appended to it. The particular combination, as described in the previous section, It is easy to see that $\mathbb{Y}^n_{x_{n+1}}(\mathcal{U})$'s corresponding contour function for Y_{n+1} is given by

$$\begin{aligned}\pi^n_{x_{n+1}}(y_{n+1}) &= \mathsf{R}_n\{\mathbb{Y}^n_{x_{n+1}}(\mathcal{U}) \ni y_{n+1}\} \\ &= \text{prob}\{\mathsf{Unif}(\mathscr{I}_{n+1}) \geq r(T_{n+1}(z^{n+1}))\} \\ &= \frac{1}{n+1}\sum_{i=1}^{n+1} 1\{T_i(z^{n+1}) \geq T_{n+1}(z^{n+1})\}.\end{aligned} \tag{29}$$

As $\mathbb{Y}^n_{x_{n+1}}(\mathcal{U})$ is both nested and non-empty, its contour function above is all that is needed to define a probabilistic predictor and, consequently, quantify uncertainty about any assertion $A \subset \mathbb{Y}$ of interest. Uniform validity of this probabilistic predictor follows directly from the general result in Theorem 1.

For illustration, consider the following example. Let $X_1, \ldots, X_n$ be iid $\mathsf{Unif}(0, 1)$, with $n = 200$, and let $Y_1, \ldots, Y_n$ be independent, where $Y_i = \mu(X_i) + 0.1\varepsilon_i$, where $\mu(x) = \sin^3(2\pi x^3)$, and $\varepsilon_1, \ldots, \varepsilon_n$ are iid from a Student-t distribution with df $= 5$. Figure 2 displays the data, the true regression function $\mu(x)$ and the fitted regression curve $\hat{\mu}(x)$ based on a B-spline with 12 degrees of freedom. A 95% prediction band is also displayed, derived by (19) and x_{n+1} taking values along the observed x^n.



(a)

(b)

Figure 2: Panel (a): Data and the plausibility contours at selected values of x. Panel (b): Data, the true mean curve (heavy line), the fitted B-spline regression curve (thin line), and the 95% pointwise prediction band.

We end this section pointing out an important connection between the prediction IM developed here and the powerful *conformal prediction* presented in Vovk et al. (2005). The reader may have recognized the Ψ function in the A-step of our construction as the so-called *non-conformity measure*, an essential component in the conformal prediction framework. Moreover, the basic output from the IM construction presented below is the plausibility contour in (29), which is precisely conformal prediction's p-value or transducer. The theory in Vovk et al. (2005) takes this conformal transducer, which is uniformly distributed as stated in Theorem 1, and constructs a prediction set as in (19) with the prediction coverage probability property as in (2). It was recently recognized (Cella and Martin 2022) that the conformal prediction output could be converted into a uniformly valid probabilistic predictor in the sense of Definition 2, one that can make valid belief assignments, by treating the transducer as the contour of a consonant plausibility function via (12). We refer to this general probabilistic predictor construction as "conformal + consonance," and all it requires is that the conformal transducer π_x^n be a plausibility contour function in the sense that it satisfy $\sup_y \pi_x^n(y) = 1$ for all (z^n, x). This is easy to verify in cases where Y is a continuous random variable. Indeed, for the Ψ function in (26), the supremum is attained at $y = \hat{\mu}^{n+1}_{-(n+1)}(x)$. In other cases, like in classification where Y is discrete, the "conformal + consonance" construction is not so straightforward. We discuss these considerations next in Section 7.

7 Probabilistic prediction in classification

In Section 6, we found that the A-step boils down to the specification of a suitable real-valued, exchangeability-preserving function Ψ, which Vovk et al. (2005) refer as a non-conformity measure. In binary classification problems, a Ψ function like in (26) can also be used here by encoding the binary labels as distinct real numbers. However, if there are more than two labels, and not in an ordinal scale where the assignment of different numbers to them is justified, there is no natural way to measure the distance between labels. Consequently, we cannot measure how wrong a prediction is—it is simply right or wrong (Shafer and Vovk 2007). To circumvent this, Vovk et al. (2005) suggest the following non-conformity measure based on nearest-neighbor classification:

$$\Psi(Z_{-i}^{n+1}, Z_i) = \frac{\min_{j \in \mathscr{I}_{n+1} \setminus \{i\} : Y_j = Y_i} d(X_j, X_i)}{\min_{j \in \mathscr{I}_{n+1} \setminus \{i\} : Y_j \neq Y_i} d(X_j, X_i)}, \tag{30}$$

where d is the Euclidean distance. In words, $\Psi(Z_{-i}^{n+1}, Z_i)$ is large if X_i is close to an element in X_{-i}^{n+1} with a label different from Y_i and far from any element in X_{-i}^{n+1} with label equal to Y_i. If both the numerator and the denominator in (30) are 0, Shafer and Vovk (2007) recommend taking the ratio also to be 0. Other non-conformity measures for classification problems can be found in Vovk et al. (2005).

Two factors were fundamental to the specification of the association (27) in Section 6, namely the identification of Ψ, so that Z^{n+1} can be mapped to T^{n+1} preserving exchangeability, and the continuity of the T_i's. In classification, however, the Y_i's are not continuous, so there could be ties in the T_i's. Consequently, their ranks would be no longer uniform distributed on $\mathscr{I}_{n+1}$. Luckily, when ties are possible, $r(T_{n+1})$ is stochastically no larger than the discrete uniform distribution it would take if there were no ties. This leads to an “association” of the form

$$r(T_{n+1}) = U, \quad U \leq_{\text{st}} \mathsf{Unif}(\mathscr{I}_{n+1}).$$

But for situations like this where the association involves a stochastic inequality, the general arguments in Martin and Liu (2015c, Sec. 5) imply that the inequality can be ignored and the association (27)—with stochastic equality—can still be used.

Having identified the appropriate association, the IM construction proceeds analogously to that in the previous section: the A-step is completed by writing (27), the random set (28) is chosen in the P-step to target the unobserved realization of the auxiliary variable U, and, in the C-step, the ingredients in the A- and P-steps are combined to get $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$, a data-dependent random subset of $\mathbb{Y}$. However, due to the discreteness of $\mathbb{Y}$, it is possible that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is empty with positive R_n-probability. As discussed in Section 5, in these cases, some adjustment to the probabilistic predictor in (22) is necessary to avoid the counter-intuitive “conflict” cases where realizations of the random set $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ happens to be empty. There is a sense in which empty prediction sets could be meaningful, but we defer this discussion to Section 8.

There are two available adjustments to account for the potentially empty realizations of the random set $\mathbb{Y}_{x_{n+1}}^n(\mathcal{S})$. The first, and probably most intuitive, is *conditioning* on the event that the random set is non-empty, which happens to be equivalent to Dempster's rule of combination (e.g., Shafer 1976, Chap. 3). For example, the post-conditioning plausibility contour is given by

$$y_{n+1} \mapsto \mathsf{R}_n\{\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) \ni y_{n+1} \mid \mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) \neq \varnothing\}.$$

It is easy to see that conditioning simply rescales the original plausibility contour, making it larger at each $y_{n+1} \in \mathbb{Y}$. Clearly, if the unadjusted probabilistic predictor is valid, then this conditioning adjustment—which only inflates its plausibility contour values—cannot fail to be valid. This inflation does, however, suggest a potential loss of efficiency, e.g., larger prediction sets in (19).

The second adjustment strategy, designed to preserve validity without sacrificing efficiency, is based on a suitable *stretching* of the original random set; see, e.g., Ermini Leaf and Liu (2012). Roughly, those $\mathcal{U}$ such that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) = \varnothing$ correspond to "conflict cases," and Dempster's conditioning rule simply removes these conflict cases and renormalizes the $\mathcal{U}$-probabilities. As an alternative, Ermini Leaf and Liu (2012) suggested to stretch those conflict $\mathcal{U}$ cases just enough so that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is non-empty. Their formulation was in the context of inference under non-trivial parameter constraints, but here we apply this to classification.

Start by defining the set

$$\mathbb{U}_{x_{n+1}}^n = \bigcup_{y_{n+1} \in \mathbb{Y}} \bigl\{ r\bigl(T_{n+1}(z^n, z_{n+1})\bigr) \bigr\} \subseteq \mathscr{I}_{n+1}. \tag{31}$$

There are only finitely many y_{n+1} values, and the set $\mathbb{U}_{x_{n+1}}^n$ defined above is just the collection of ranks that are possible for the given Z^n and x_{n+1}. Note that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is empty if and only if $\mathcal{U}$ has empty intersection with $\mathbb{U}_{x_{n+1}}^n$. Therefore, the conflict cases mentioned above can be alternatively defined as realizations of $\mathcal{U}$ that have empty intersection with $\mathbb{U}_{x_{n+1}}^n$. This conflicting situation can be avoided if, instead of throwing out the conflict $\mathcal{U}$, we stretch it to a suitable $\mathcal{U}_e$, with $e \geq 0$ a stretching parameter that controls how far $\mathcal{U}$ is stretched toward $\mathbb{U}_{x_{n+1}}^n$. In particular, we take

$$\mathcal{U}_e = \{1, 2, \ldots, U' + e\}, \quad U' \sim \mathsf{Unif}(\mathscr{I}_{n+1}).$$

Following Ermini Leaf and Liu (2012), the parameter e is chosen as the smallest value at which the intersection of $\mathcal{U}_e$ and $\mathbb{U}_{x_{n+1}}^n$ is non-empty, i.e.,

$$\hat{e} = \min\{e : \mathcal{U}_e \cap \mathbb{U}_{x_{n+1}}^n \neq \varnothing\} = \begin{cases} \min \mathbb{U}_{x_{n+1}}^n - U' & \text{if } U' < \min \mathbb{U}_{x_{n+1}}^n \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\mathcal{U}_{\hat{e}}$ would be

$$\mathcal{U}_{\hat{e}} = \begin{cases} \{1, 2, \ldots, \min \mathbb{U}_{x_{n+1}}^n\} & \text{if } U' < \min \mathbb{U}_{x_{n+1}}^n \\ \{1, 2, \ldots, U'\} & \text{otherwise.} \end{cases}$$

In summary, in the stretching IM, the IM's original random set output $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is replaced with $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}_{\hat{e}})$, and its guaranteed non-emptiness makes the probabilistic predictor derived from it valid. It is also more efficient than conditioning since it avoids globally inflating the plausibility contour via renormalization, as the following example highlights.

For illustration, consider the data in Table 1, taken from Agresti (2003, p. 304), describing the primary food choices and lengths of $n = 39$ male alligators caught in Lake George, Florida. Assume the 40th caught alligator is two meters long, i.e., $X_{n+1} = 2$. The goal is to predict Y_{n+1}, its primary food choice. Note that

$$\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) = \begin{cases} \{I\} & \text{with probability } 0.1 \\ \{I, F\} & \text{with probability } 0.2 \\ \{I, F, O\} & \text{with probability } 0.3 \\ \varnothing & \text{with probability } 0.4. \end{cases} \tag{32}$$

Length (m)	Choice	Length (m)	Choice	Length (m)	Choice
1.30	I	1.65	I	2.03	F
1.32	F	1.65	F	2.31	F
1.32	F	1.68	F	2.36	F
1.40	F	1.70	I	2.46	F
1.42	I	1.73	O	3.25	O
1.42	F	1.78	F	3.28	O
1.47	I	1.78	O	3.33	F
1.47	F	1.80	F	3.56	F
1.50	I	1.85	F	3.58	F
1.52	I	1.93	I	3.66	F
1.63	I	1.93	F	3.68	O
1.65	O	1.98	I	3.71	F
1.65	O	2.03	F	3.89	F

Table 1: Primary food choice (I, invertebrates; F, fish; O, other) and lengths (in meters) for $n = 39$ male alligators (Agresti 2003, p. 304).

The corresponding plausibility contour, as given in (23), is represented by the solid lines in Figure 3(a). By thresholding it at any $\alpha > 0.6$ we obtain $100(1-\alpha)\%$ prediction sets that are empty, which is undesirable.

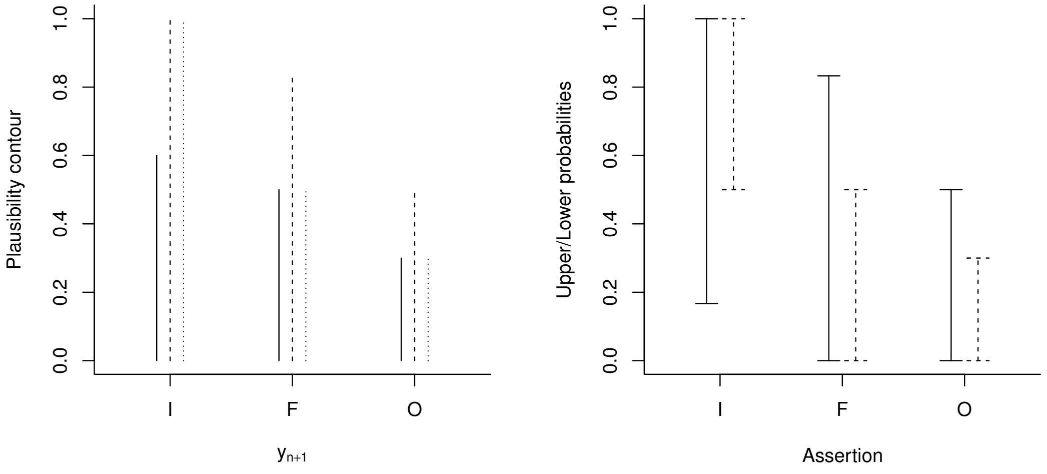
The plausibility contour conditioned on (32) $\neq \varnothing$ is easy to evaluate, and is represented by the dashed lines in Figure 3(a). To calculate the plausibility contour under the stretching approach, we obtain, after some calculations, $\mathbb{U}^n_{x_{n+1}} = \{17, 21, 29\}$. As $\min \mathbb{U}^n_{x_{n+1}} = 17$,

$$\mathcal{U}_{\hat{e}} = \begin{cases} \{1, 2, \ldots, 17\} & \text{if } U' < 17 \\ \{1, 2, \ldots, U'\} & \text{otherwise.} \end{cases}$$

where $U' \sim \mathsf{Unif}(1, 2, \ldots, 40)$. Therefore,

$$\mathbb{Y}^n_{x_{n+1}}(\mathcal{U}_{\hat{e}}) = \begin{cases} \{I\} & \text{with probability } 0.5 \\ \{I, F\} & \text{with probability } 0.2 \\ \{I, F, O\} & \text{with probability } 0.3, \end{cases}$$

and the dotted lines in Figure 3(a) illustrate its corresponding plausibility contour. Note, first, that empty prediction sets are eliminated with both the conditioning and the stretching adjustments. Second, for any α, the $100(1-\alpha)\%$ prediction sets derived from the stretching adjustment are no larger than the corresponding ones derived from the conditioning adjustment, which indicates that the former is no less efficient than the latter. Another way to see this is through the difference between the upper and lower probabilities derived by the respective probabilistic predictors. Dempster (2008) referred to this gap as the "don't know" probability. Of course, between two valid probabilistic predictors, the one with less "don't know" is preferred because it is more efficient. Figure 3(b) shows the upper and lower probabilities for the singleton assertions $\{I\}$, $\{O\}$ and $\{F\}$, for both strategies. Clearly, stretching leads to a more efficient probabilistic predictor.



(a) (b)

Figure 3: Panel (a): Plausibility contours in Equation (23), derived from an IM construction with no adjustment (solid lines), conditioning adjustment (dashed lines) and stretching adjustment (dotted lines). Panel (b): Upper and lower probabilities for the singleton assertions $\{I\}$, $\{F\}$ and $\{O\}$ derived from an IM construction with the conditioning adjustment (solid lines) and the stretching adjustment (dashed lines). These predictions are based on a new alligator of length $x_{n+1} = 2$ meters.

To further see this gain in efficiency we consider the *Glass Identification* data set from the USA Forensic Science Service, available in the UCI Machine Learning Repository (Dua and Graff 2017).[1] It has 10 attributes associated with 214 glasses. The type of glass, a categorical variable—with six categories, including "containers" and "headlamps"—is the response variable. The nine remaining variables, which describe the oxide content, i.e., Na, Fe, K, etc., are the explanatory variables. Classification of types of glass is relevant in criminology applications, where glass fragments left at the scene of the crime may be important evidence if correctly identified. To evaluate the performance in classifying glass fragments, we randomly split the data in half and train both the conditioning and stretching strategies in the first half, with Ψ function as in (30). For further comparison, we also train a Bayesian multinomial regression model with default, non-informative priors on the parameters.[2] Figure 4 plots the distribution functions of the corresponding plausibility contours for the responses in the second half of the data. As expected, uniform validity in (8) fails for the Bayesian solution and holds for both IM solutions, with the one based on stretching being more efficient. Of course, not being uniformly valid does not imply that the Bayesian prediction set will not achieve the nominal coverage, but we can check this directly. Table 2 shows the empirical coverage probabilities and the average sizes (cardinality) of 95% prediction sets for the responses in the second half of

[1] `https://archive.ics.uci.edu/ml/datasets/glass+identification`

[2] The *bamlss* R package (Umlauf et al. 2021) was used to run the Bayesian analysis.

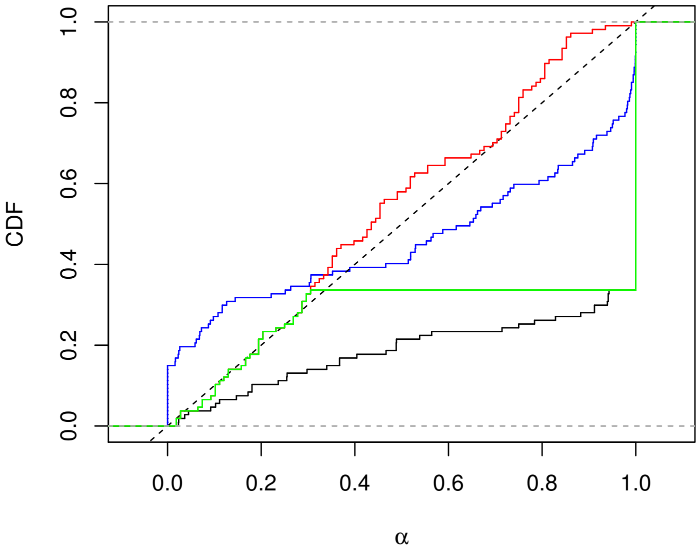


Figure 4: Distribution function for plausibility contours derived from an IM construction, with conditioning (black), stretching (green) and no adjustment (red), and a Bayesian multinomial regression model (blue).

Strategy	Coverage	Size
Conditioning	0.96	3.07
Stretching	0.96	2.73
Bayes	0.80	1.71

Table 2: Coverage probabilities and average size of 95% prediction sets in (19) derived from an IM construction, with conditioning and stretching adjustment, and a Bayesian multinomial regression model.

the data. Clearly, the Bayes approach does not provide valid prediction sets.

Recall from Section 6 that the probabilistic predictor derived from the "conformal + consonance" construction is uniformly valid according to Definition 2, given that the conformal transducer π_x^n satisfies (11). In regression problems, this condition follows naturally from the continuity of Y, and the derived probabilistic predictor is equivalent to the one that would be obtained from an IM construction (assuming both use the same Ψ function). In classification problems, however, (11) may not hold because Y is discrete. This implies the "conformal + consonance" cannot be applied directly without some adjustment. This is not surprising given that similar adjustments were needed in the IM construction discussed above too. To better see this, note that the distribution function for the conformal transducer is also shown in Figure 4. The need of an adjustment is evident, as uniform validity fails and, consequently, the derived conformal prediction intervals obtained through (19) would not be calibrated for certain choices of α.

A natural adjustment is to force the conformal transducer to attain the value 1.

Consider the following two adjusted conformal transducers:

$$\dot{\pi}_x^n(y) = \frac{\pi_x^n(y)}{\max_y \pi_x^n(y)},$$

and

$$\ddot{\pi}_x^n(y) = \begin{cases} 1 & \text{if } y = \hat{y}, \\ \pi_x^n(y) & \text{otherwise,} \end{cases}$$

where $\hat{y} = \arg\max_y \pi_x^n(y)$ and $y \in \mathbb{Y}$. In words, $\dot{\pi}_x^n(y)$ takes the conformal transducers for the different $y \in \mathbb{Y}$ and divide them by their maximum, and $\ddot{\pi}_x^n(y)$ maintains all the conformal transducer values except for its maximum, which is assigned the value 1. That both adjusted transducers reach the value 1 makes the probabilistic predictors derived by them, through (12), uniformly valid in the sense of Definition 2. It is also easy to see that these probabilistic predictors obtained from $\dot{\pi}_x^n(y)$ and $\ddot{\pi}_x^n(y)$ are equivalent to the ones derived from the IM construction with, respectively, the conditioning and the stretching adjustments. This shows that forcing consonance of the conformal transducer is not an ad hoc strategy; it is justified by the corresponding operations on random sets. Moreover, in light of this connection to the IM's random set adjustments, we find that the second adjustment to the conformal predictor, i.e., setting the maximum value equal to 1, is the more efficient adjustment.

8 Conclusion

Here we focused on the important problem of prediction in supervised learning applications with no model assumptions (except exchangeability). We presented a notion of prediction validity, one that goes beyond the usual coverage probability guarantees of prediction sets. This condition assures the reliability of the degrees of belief, obtained from a imprecise probability distribution, assigned to all relevant assertions about the yet-to-be-observed quantity of interest. We also showed that, by following a new variation on the (generalized) IM construction first presented in Martin (2015, 2018), this validity property can be easily achieved. We also noted the connection between this new IM construction and the conformal prediction strategy in, e.g., Vovk et al. (2005), and presented illustrations in both regression and classification settings. This connection is of paramount importance, as it implies that no new methodology is needed to achieve the (uniform) validity properties presented here. All that is needed is a possibilistic interpretation of the conformal prediction output.

Exchangeability was crucial to our IM construction, that is, without exchangeability, we cannot establish the distribution of the auxiliary variables. While exchangeability is a relatively weak assumption compared to iid from a parametric family, there are, of course, situations where exchangeability is inappropriate, such as time series or spatial applications. Work to develop conformal prediction methods in not-exactly-exchangeable settings is an active area of current research (e.g., Mao et al. 2020), and it would be interesting to see what the IM perspective has to offer here.

In Section 5 we noted that the IM construction there leads naturally to a notion of *marginal* validity, which is different (and weaker) than the so-called *conditional* validity property. While this is usually framed in the context of prediction sets, the corresponding definition in the context of probabilistic predictors is

$$\mathsf{P}\{\overline{\Pi}_x^n(A) \leq \alpha, Y_{n+1} \in A \mid X_{n+1} = x\} \leq \alpha \quad \forall\, x,$$

and, of course, for all $(\alpha, n, A, \mathsf{P})$ as before. Given the impossibility results in, e.g., Lei and Wasserman (2014), it seems unlikely that conditional validity can be achieved by any non-trivial probabilistic predictor. Asymptotic conditional validity is possible, and some promising ideas are given in, e.g., Chernozhukov et al. (2019).

We mentioned in Section 7 that, surprisingly, empty random sets may have some practical value. This concerns the so-called *open-* versus *closed-world* view of the prediction problem. If the world is closed in the sense that all the possible labels are known, then it makes sense to remove the empty set cases and, hence, force consonance. However, if the world is open in the sense that other labels are possible, then the empty set realization is an indication that the new object being classified may be of previously-unknown type, which itself is valuable information. How this open-world view can be captured by the IM framework developed here remains an open question.

Acknowledgments

The authors thank the reviewers from the conference proceedings and journal submissions for their valuable feedback, and the *IJAR* guest editors—Andrés Cano, Jasper De Bock, and Enrique Miranda—for the invitation to contribute to the special journal issue. This work is partially supported by the U.S. National Science Foundation, grants DMS–1811802 and SES–2051225.

References

Agresti, A. (2003). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):1–20.

Cahoon, J. and Martin, R. (2020). Generalized inferential models for meta-analyses based on few studies. *Statistics and Applications*, 18(2):299–316.

Cahoon, J. and Martin, R. (2021). Generalized inferential models for censored data. *International Journal of Approximate Reasoning*, 137:51–66.

Campi, M., Calafiore, G., and Garatti, S. (2009). Interval predictor models: Identification and reliability. *Automatica*, 45(2):382–392.

Cella, L. and Martin, R. (2021a). Approximately valid and model-free possibilistic inference. In Denœux, T., Lefèvre, E., Liu, Z., and Pichon, F., editors, *Belief Functions: Theory and Applications*, pages 127–136, Cham. Springer International Publishing.

Cella, L. and Martin, R. (2021b). Valid inferential models for prediction in supervised learning problems. In Cano, A., De Bock, J., Miranda, E., and Moral, S., editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, pages 72–82. PMLR.

Cella, L. and Martin, R. (2022). Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, 141:110–130.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2019). Distributional conformal prediction. `arXiv:1909.07889`.

Coolen, F. P. A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language, and Information*, 15(1/2):21–47.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statististics*, 38:325–339.

Dempster, A. P. (1968). A generalization of Bayesian inference. (With discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 30:205–247.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377.

Dempster, A. P. (2014). Statistical inference from a Dempster–Shafer perspective. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press.

Denœux, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252.

Denœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547.

Denœux, T. and Li, S. (2018). Frequency-calibrated belief functions: review and new insights. *International Journal of Approximate Reasoning*, 92:232–254.

Dua, D. and Graff, C. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. University of California, Irvine, School of Information and Computer Sciences.

Dubois, D. and Prade, H. (1988). *Possibility Theory.* Plenum Press, New York.

Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *International Journal of Approximate Reasoning*, 53(5):709–727.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference.* Hafner Press, New York, 3rd edition.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* Springer-Verlag, New York.

Gong, R. and Meng, X.-L. (2021). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statistical Science*, 36(2):169–190.

Grünwald, P. (2018). Safe probability. *Journal of Statistical Planning and Inference*, 195:47–63.

Hahn, P. R., Martin, R., and Walker, S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, 113(523):1085–1093.

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. `arXiv:1807.00263`.

Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Mao, H., Martin, R., and Reich, B. (2020). Valid model-free spatial prediction. `arXiv:2006.15640`.

Martin, R. (2015). Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, 110(512):1552–1561.

Martin, R. (2018). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, 195:105–115.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73.

Martin, R. (2021). An imprecise-probabilistic characterization of frequentist statistical inference. *Researchers.One*, `https://researchers.one/articles/21.01.00002`.

Martin, R. and Lingham, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics*, 58:225–235.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108:301–313.

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:195–217.

Martin, R. and Liu, C. (2015b). *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.

Martin, R. and Liu, C. (2015c). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *Journal of the American Statistical Association*, 110(512):1621–1631.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shafer, G. and Vovk, V. (2007). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Umlauf, N., Klein, N., Simon, T., and Zeileis, A. (2021). bamlss: A Lego toolbox for flexible Bayesian regression (and beyond). *Journal of Statistical Software*, 100(4):1–53.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Vovk, V. and Petej, I. (2014). Venn–abers predictors. In Zhang, N. L. and Tian, J., editors, *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pages 829–838. AUAI Press.

Vovk, V., Shen, J., Manokhin, V., and Xie, M. (2018). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *Journal of Statistical Planning and Inference*, 105:35–65.

Wang, C. M., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142(7):1980–1990.