Journal of Machine Learning Research 27 (2026) 1-28 Submitted 7/23; Revised 5/26; Published 6/26

Model-free generalized fiducial inference

Jonathan P Williams JWILLI27@NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC, USA

Editor: Manuel Gomez-Rodriguez

Abstract

Conformal prediction (CP) was developed to provide finite-sample probabilistic prediction guarantees. While CP algorithms are a relatively general purpose approach to uncertainty quantification, with finite-sample guarantees, they lack versatility. Namely, the CP approach does not *prescribe* how to quantify the degree to which a data set provides evidence in support of (or against) an arbitrary event from a general class of events. This paper uses tools from imprecise probability theory to build a formal connection between CP and generalized fiducial (GF) inference. These new insights establish a more general inferential lens from which CP can be understood, and demonstrate the pragmatism of fiducial ideas. The formal connection establishes a context in which epistemically-derived GF probability matches aleatoric/frequentist probability. Beyond this fact, this work illustrates how tools from imprecise probability theory, namely lower and upper probability functions, can be applied in the context of the imprecise GF distribution to provide posterior-like, prescriptive inference that is not possible within the CP framework alone. In addition to the primary CP generalization that is contributed, fundamental connections are synthesized between this new model-free GF and three other areas of contemporary research: nonparametric predictive inference (NPI), conformal predictive systems/distributions, and inferential models (IMs).

Keywords: approximate reasoning, Dempster-Shafer theory, foundations of statistics, possibility theory, safe probability

1. Introduction

The rate at which machine learning algorithms are being developed and advanced for high-stakes applications is greatly exceeding the pace at which safe and reliable methods for quantifying their uncertainty are becoming understood. Moreover, it is largely undetermined how to properly quantify uncertainty in their performance guarantees, and there is no generally accepted standard of accountability of stated uncertainties. As discussed in Shafer (2021)—among many other references—a trouble with non-frequentist interpretations of probability is their practical limitations for verifiability. Frequentist interpretations of probability yield explicit definitions based on probabilistic statements that can be tested and verified (if only through theoretical simulation), and admit tangible attributes of data models, such as *validity* of predictions (e.g., control over type 1 error rates). Notions of validity are fundamental to developing methods and procedures that have any chance at being reliable when applied in the context of uncertainty quantification (for prediction and

inference, alike). It is for these reasons that statisticians must afford a high premium to repeated sampling properties. CP was developed to provide finite-sample probabilistic prediction guarantees by leveraging the calibration inherent in applying the empirical hold-out method for training/testing a machine learning prediction rule (Vovk et al., 2005). Growing momentum for applications and developments of CP has occurred in recent years.

While CP algorithms are a relatively general purpose approach to uncertainty quantification, with finite-sample guarantees, they lack versatility; whereas Bayesian approaches, for instance, are versatile but lacking in guarantees. Namely, the CP approach does not *prescribe* how to quantify the degree to which a data set provides evidence in support of (or against) an arbitrary event from a general class of events. In contrast, within the Bayesian paradigm, the degree to which a data set provides evidence in support of (or against) an event is quantified by the posterior probability of the event, *for any measurable event*. Bayesian inference, however, operates by the usual Kolmogorov axioms for probability calculus, and is thereby subject to the false confidence theorem (Balch et al., 2019; Martin, 2019; Carmichael and Williams, 2018), rendering it provably unreliable. The false confidence theorem is mathematical justification for the fact that precise probabilistic-based statistical inferences (e.g., those based on posterior probabilities) are provably unreliable in the sense that there always exists a false hypothesis (with positive Lebesgue measure) having arbitrarily large epistemic (e.g., posterior) probability, with arbitrarily large aleatory (i.e., frequency/frequentist) probability. This theorem arose to explain a troubling phenomenon occurring in the Bayesian analysis of satellite trajectory data (Balch et al., 2019). Consequences of the false confidence theorem can, however, be avoided via imprecise probability calculus (Martin, 2019).

This paper uses tools from imprecise probability theory to build a formal connection between CP and GF inference (Hannig et al., 2016). These new insights establish a more general inferential lens from which CP can be understood, and demonstrate the pragmatism of fiducial ideas. The key observation about the CP framework is that the rank of a nonconformity score actually defines a *data generating association* with an auxiliary discrete-uniform distribution. This article proves that applying the GF inference framework to this rank-based data generating association leads to a model-free approach—hereafter referred to as *model-free GF*—for constructing GF prediction sets, from which CP sets arise. This also means that the epistemically-derived GF probability matches the aleatoric/frequentist probability on CP sets. Beyond this fact, this work illustrates how tools from imprecise probability theory, namely lower and upper probability functions, can be applied in the context of the imprecise GF distribution to provide posterior-like, prescriptive inference that is not possible within the CP framework alone. Furthermore, beyond the primary contribution on CP generalization, this paper synthesizes fundamental connections between this new model-free GF and three other areas of contemporary statistics research: NPI, conformal predictive systems/distributions, and IMs.

First, the model-free GF framework turns out to be a generalization of the NPI approach (e.g., Coolen, 1998; Augustin and Coolen, 2004; Coolen and Coolen-Maturi, 2024) built from the Dempster-Hill procedure/assumption (Dempster, 1963; Hill, 1968, 1988). At a high level, NPI is a statistical methodology that assigns uniform probability $1/(n+1)$ to each interval between the order statistics of a real-valued data set of size n, and yields direct conditional probabilities about future observable random quantities (Coolen and

Coolen-Maturi, 2024). Within the model-free GF paradigm, the Dempster-Hill assumption is satisfied trivially and more generally, and under this assumption, non-asymptotic, sub-exponential concentration inequalities are derived to establish root-n consistency, around the true distribution of the data, of every probability measure in the credal set of the imprecise model-free GF distribution.

Second, conformal predictive distributions (e.g., Vovk et al., 2019; Vovk, 2022, 2024) are also an extension of NPI ideas but with the aim of constructing predictive probability distribution functions from monotonic nonconformity scores. Conformal predictive distributions correspond to cumulative distribution functions that yield a finite-sample coverage guarantee under a nonparametric assumption; but unlike usual CP not all nonconformity measures will produce a conformal predictive distribution (Vovk et al., 2019). This approach is advertised as fiducial in the sense of constructing a probability distribution from a pivot, but is more aligned with the confidence distribution characterization of fiducial ideas (i.e., Xie and Singh, 2013; Schweder and Hjort, 2016) rather than the GF or IM characterizations. In contrast, the model-free GF characterization does not rely on monotonicity of nonconformity scores, nor does it lead to a precise probabilistic predictive distribution. In fact, the model-free GF imprecise probabilistic characterization is more aligned with the NPI prescriptions for inference, while still leveraging CP-calibrated constructions beyond the Dempster-Hill procedure/assumption.

Third, IM is an alternative characterization of the fiducial principles/approach that Fisher introduced in the early twentieth century (Fisher, 1930, 1935), but one that is fully prescriptive, frequentist-calibrated, and built from a reorientation of the epistemic interpretation of the imprecise probability tools developed from DS calculus. The data generating association implemented to construct the model-free GF distribution by way of the GF *switching principle* (Hannig et al., 2016) can alternatively be implemented by way of a *predictive random set* to construct an IM (Martin and Liu, 2015) for predictive inference, as in Cella and Martin (2022). Similar to the model-free GF construction, this IM construction inherently elicits the formulation of CP sets. The difference between the two constructions is that model-free GF has belief and plausibility functions corresponding to a basic probability assignment function (using the language of Dempster-Shafer (DS) theory; see Definition 9 below) that assigns nonzero mass only to the disjoint set-differences of the nested CP sets, whereas the IM lower and upper probabilities are more directly associated with the CP sets. It turns out that the disjoint nature of the focal sets, as in the model-free GF construction, is necessary for the statistical consistency of the probability measures in the credal set. The IM predictive distribution in Cella and Martin (2022) was designed instead to achieve an imprecise probabilistic notion of finite-sample frequentist validity, a property endowed by *consonance* (Shafer, 1976; Cella and Martin, 2022); the consonant IM construction is similarly established using the model-free GF framework in Section 4.2, thereby also establishing imprecise probabilistic validity.

1.1 Related work

Since precise probabilistic approximations from the credal set associated with model-free GF belief/plausibility functions may be desirable, it is illustrated in this article how one such approximation can be arrived at based on the principle of maximum entropy. This should

be consistent with the pignistic transformation based on the generalized insufficient reason principle advocated for in Smets (1990). A mapping from a credal set to a probability distribution is often referred to as *a pignistic transformation*; the terminology being that beliefs are held at the credal level while decisions are made at the pignistic level (Smets, 1990). The problem of constructing pignistic transformations has been considered, more broadly, in the approximate reasoning research community. For example, Dubois et al. (2004) provides conditions and theorems for transforming between possibility/necessity measures and probability measures. Still, no general theory for such constructions or evaluations of such mappings exists (Grünwald, 2018), and so there remain many open research questions concerning precise probability approximations of a credal set, especially from a statistical inference perspective.

Imprecise probabilities, and in particular, the roles of [non-additive] belief and plausibility functions have been extensively developed within the context of DS theory (Dempster, 1966; Shafer, 1976). DS theory has seen varied applications, namely in artificial intelligence communities (e.g., Bloch, 1996; Vasseur et al., 1999; Denoeux, 2000; Basir and Yuan, 2007; Denoeux, 2008; Díaz-Más et al., 2010), but has largely not been applied in mainstream statistical literatures. The lack of attention from the statistics communities may be partly attributed to major barriers to computation (Shafer, 2021). Remarkably, a recent solution drawing positive attention has been provided in the article Jacob et al. (2021) for the computation of DS inference on categorical data, a problem that had been open for 55 years. Regardless of the success/failure of the DS theory for inference, the related ideas developed for lower and upper probabilities in Shafer (1976) are very useful and apply more broadly, as demonstrated with the IM framework. In particular, the utility of a *don't know* category has hugely important implications on statistical inference, as demonstrated/discussed in Balch et al. (2019); Martin (2019); Carmichael and Williams (2018); Williams (2021).

Another line of work related to the model-free GF formulation of CP sets is the development of *risk controlling prediction sets* (Bates et al., 2021; Angelopoulos et al., 2023, 2025). In the risk controlling prediction set framework, CP ideas are generalized to provide guarantees on the expected loss between the response label and CP set, where the specification of the loss function is arbitrary (based on minimal assumptions). In principle, the model-free GF ideas could also be applied to exchangeable instances of loss functions, but the imprecision of the model-free GF formulation more naturally lends itself to the analysis of *upper/lower previsions* of a loss (i.e., supremum/infimum of the expectation of the loss function, over a credal set of probability measures). Indeed, the risk controlling prediction set perspective inspired investigation of the latter in Williams and Liu (2024).

The remainder of this paper is organized as follows. Section 2 lays out a general discussion and motivating remarks on various notions of safety and reliability in prediction and inference; namely, such notions go far beyond the unconditional validity emphasized in this article. Section 3 serves to introduce the fundamental ideas for CP and the GF inference paradigm, followed by construction of the framework being proposed for *model-free GF inference*. The imprecise nature of this construction is examined in Section 4, including connections to NPI, conformal predictive distributions, and IMs. An illustration is offered in Section 5 for the construction of a pignistic transformation from the model-free GF credal set to a [precise] probability distribution, followed by a brief survey of other such transformations implicitly considered in the GF literature. Concluding remarks are provided in Section

6, and the Julia programming language codes to reproduce all numerical illustrations and figures are publicly available at: `https://jonathanpw.github.io/research.html`.

2. Safe probability, testing, and inference

The developments in this article are governed by a particular notion of reliability, namely finite-sample, unconditional, type 1 control of prediction error; commonly this is referred to as a type of *unconditional validity*. Consider the following motivating remarks.

At the American Society of Clinical Oncology conference in Chicago in June 2022 (Tie et al., 2022), it was discussed that a new liquid biopsy can help identify the need for adjuvant therapy in stage II colon cancer, thereby potentially avoiding post-operative chemotherapy, which, for colon cancer, can cause peripheral neuropathy. Suppose a machine learning algorithm is trained to identify the need for adjuvant therapy with 95% confidence reported. How is this confidence defined? Is it defined as the reported error on a test set? Is it a Bayesian posterior probability? Is it some sort of averaging over a collection of predictions? All of these are widely accepted notions of *confidence*, but they all represent different quantifications of uncertainty with varying (if any) guarantees for how the algorithm might perform on future data. When a meteorologist predicts that there is 95% chance of rain tomorrow, it might not be so problematic to not understand in what sense *95% chance* is reliable or verifiable (if at all), but when an algorithm says there is 95% chance that a person does not need post-operative chemotherapy with potentially life-debilitating side effects, there are serious ramifications for how to interpret that quantification of uncertainty. Moreover, for such quantification of uncertainty to be reliable there must be a mechanism for accountability. Finite-sample, unconditional, type 1 error control facilitates such accountability in that 95% confidence is defined unambiguously by the algorithmic attribute that in no more than 5% of instances of rejecting the hypothesis 'post-operative chemotherapy is needed' will post-operative chemotherapy actually be needed—an error rate that can be verified on real data of any sample size.

CP and model-free GF are reliable in this sense of unconditional validity for uncertainty quantification of predictions. For statistical inference of a population parameter, IMs and, more recently, universal inference (Wasserman et al., 2020; Park et al., 2026) guarantee the same type of unconditional validity, as does generalized universal inference (Dey et al., 2025) for a risk minimizer; see also Dey and Williams (2023) on valid inference for machine learning model parameters. Unconditional validity has limitations, though. For instance, suppose that of colon cancer patients, 99% are of age 45 years or older, and that the type 1 error rate of rejecting 'post-operative chemotherapy is needed' is 4% for such patients but 100% for patients younger than 45. Unconditionally, the error rate is 4.96% in this example, so the algorithm is reliable with respect to a stated guarantee of 5% error rate, but it is clearly not safe for young colon cancer patients. A stronger notion of *conditional validity*, e.g., finite-sample type 1 error control conditioned on covariate values, is required, but not readily achievable within the CP or model-free GF constructions, unless further information is available. Note that this is more general than the special case of all meaningful values of a covariate being known a priori, in which case Mondrian CP (Vovk et al., 2005; Hjort et al., 2025, and analogously model-free GF) could be applied (essentially training independently on each covariate-valued class).

Beyond unconditional and conditional validity, the article Grünwald (2018) builds an entire theory of various notions of reliability, therein termed "safety" and hence the title of this section. Loosely, in the order of strength of the guarantees offered, these notions are: asymptotic validity, unbiasedness, confidence safety, pivotal safety, calibration, squared-error optimality, unconditional validity, conditional validity. Calibration of an algorithm means that an event actually occurs in $p \cdot 100\%$ of observed instances, for events that are predicted to have probability p by the algorithm, e.g., rainfall occurring on approximately 30% of days for which the algorithm predicted 'probability of rain is 0.3'. More recently, notions of reliability have been introduced and developed in the context of sequential testing problems, termed *anytime-validity*; i.e., reliability in the sense of finite-sample type 1 error rate control under *optional continuation* or *optional stopping* of statistical testing regimes determined before or after data are observed. These properties are achieved via the construction of *e-values* and *e-processes*, by leveraging Markov's inequality and properties of non-negative supermartingales such as Ville's inequality. Furthermore, whereas p-values generally cannot be combined (e.g., across studies) to produce a p-value, e-values can; e.g., the average of e-values is an e-value. See the following references for a current review on anytime-validity and related topics: Wasserman et al. (2020); Ramdas et al. (2023); Grünwald et al. (2024); Grünwald (2024).

3. Constructing CP sets from GF inference

Throughout this text the notion of a population parameter will be used to refer to unknown population quantities of interest. Depending on the context, a parameter may take an arbitrary value. Most common examples of parameters are objects described in, but not limited to, scalar-, vector-, or matrix-valued form; though, a population parameter of interest may also be defined as an infinite-dimensional object such as a distribution function, for example. In the case of prediction, the unknown parameter value is the datum value to be predicted. For a random sample $y_1, \ldots, y_n$, of size n, denote y_{n+1} as the datum value to be predicted, and assume that these values are, respectively, realizations of the random variables $Y_1, \ldots, Y_n, Y_{n+1} \overset{\text{iid}}{\sim} Y$, where Y represents the random variable from a population model. Moving forward, the shorthand $y \sim Y$ is taken to mean y is an observed instance of the random variable Y.

Classical statistical inference on the unknown value of y_{n+1} would be to assume a parametric model for Y and construct prediction sets either inspired by large-sample theory (e.g., an asymptotic confidence interval) or from a Bayesian posterior predictive distribution. While both approaches are considered reasonable, they both allow the practitioner to avoid accountability to a stated nominal level of confidence. Without knowledge of the population model, the parametric prediction sets are not guaranteed to achieve their nominal coverage (at least non-asymptotically), i.e., a purported $1-\alpha$ level set, $\alpha \in (0,1)$, may not contain the realized value of Y_{n+1} in at least $(1-\alpha)100\%$ of repeated samples; and Bayesian posterior credible sets are not promised to be calibrated to any notion of reliability. This is highly problematic because without rigorous justification of a stated nominal level of confidence, practitioners can claim any level without consequence, and so it is not clear in what sense *probabilities* are meaningful. We need probabilities assigned to prediction sets to be inherently meaningful in a manner that is mathematically or empirically verifiable, and

so we must begin by defining the properties they ought to have. Such is the fundamental principle of the CP approach, discussed next. Note that finite-sample coverage of prediction sets at stated levels is synonymous with finite-sample type 1 error rate control at stated levels, attributable to a mathematical duality between these two notions of *validity*.

For any a priori fixed $\alpha \in (0,1)$, suppose that Γ_n^α is an α level prediction set for y_{n+1}, constructed from observed data $y_1, \ldots, y_n$. Next, assuming $y_{n+1} \sim Y$, let ξ be the binary indicator of the event that Γ_n^α does *not* contain y_{n+1}, and take $\xi_1, \xi_2, \ldots$ to represent a sequence of independent, repeated samples of ξ. Then Γ_n^α is said to be (conservatively) *valid* if $\xi_1, \xi_2, \ldots$ is dominated in distribution by that of a sequence of iid α weighted coin tosses (Vovk et al., 2005). This notion is stated more concisely in Definition 1.

Definition 1 (Type 1 validity – Cella and Martin, 2022) *Let $\{\Gamma_n^\alpha : \alpha \in (0,1)\}$ be a family of prediction sets constructed from observed data $y_1, \ldots, y_n \sim Y$, and assume $y_{n+1} \sim Y$. Denoting by P the probability measure associated with Y, the family of prediction sets is type 1 valid if, for all (α, n, P), $P\big(\Gamma_n^\alpha \ni Y_{n+1}\big) \geq 1 - \alpha$.*

Attributable to an emphasis on bounding type 1 errors, conservative validity is often simply referred to as validity. It turns out that CP sets are valid in this sense, for finite sample sizes, as discussed next.

3.1 Conformal prediction

The basic principle for any CP set is that it is constructed from an algorithm providing finite-sample guarantees, called a *conformal algorithm* and stated here as Algorithm 3. Perhaps the simplest context for introducing a conformal algorithm is the classification scenario where *exchangeable* examples $y_1, \ldots, y_n \sim Y$ are observed, as in Definition 2, and the task is to determine whether some new value y is exchangeable with $y_1, \ldots, y_n$. Note that exchangeability of data is a slightly weaker condition than assuming iid data.

The CP strategy is to first define a measure of *nonconformity*, $\Psi : \bigcup_n \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, such that $\Psi(y_{-i}^{n+1}, y_i)$, for $i \in \{1, \ldots, n+1\}$, is a meaningful measure of how different the value y_i is from the values $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{n+1}$, where $y_{-i}^{n+1} := \{y_1, \ldots, y_{n+1}\} \setminus \{y_i\}$. Then the assertion that y_{n+1} is exchangeable with $y_1, \ldots, y_n$ is dismissed if the value $\Psi(y_{-(n+1)}^{n+1}, y_{n+1})$ falls in the α tail region of the empirical distribution of the values $\Psi(y_{-i}^{n+1}, y_i)$, for $i \in \{1, \ldots, n+1\}$. When the context is clear, for conciseness let $t_i(y_i) := \Psi(y_{-i}^{n+1}, y_i)$, for $i \in \{1, \ldots, n+1\}$.

Using the conformal algorithm, a CP set denoted by Γ_n^α is constructed as the set of all y such that the conformal algorithm returns value 1. As exhibited by Theorem 4, the novelty of the conformal algorithm is its finite-sample control of type 1 errors at the stated nominal level α, for any user-specified level $\alpha \in (0,1)$, and it is sufficient to only assume exchangeability of the data examples (i.e., a model-free assumption). In fact, the exchangeability of the data is not necessary so long as $t_1(Y_1), \ldots, t_{n+1}(Y_{n+1})$ are exchangeable.

Definition 2 (Exchangeability) *A sequence $Y_1, Y_2, \ldots$ with probability measure P is said to be exchangeable if for every integer $n > 0$, every permutation σ on $\{1, \ldots, n\}$, and every P-measurable set E, $P\big\{(Y_1, \ldots, Y_n) \in E\big\} = P\big\{(Y_{\sigma(1)}, \ldots, Y_{\sigma(n)}) \in E\big\}$.*

Algorithm 3: Conformal algorithm (Vovk et al., 2005).

Input: Nonconformity measure $\Psi : \big(\mathbb{R}^n \big) \times \mathbb{R} \to \mathbb{R}$, measurable; exchangeable examples $y_1, \ldots, y_n$; an arbitrary value y; and significance level $\alpha \in (0, 1)$.

Output: Logical value; 1 indicates that $y_1, \ldots, y_n, y$ are exchangeable, and 0 else.

```
Denote y_{n+1} := y;
for i ∈ {1, ..., n+1} do
    Compute t_i(y_i) = Ψ(y_{-i}^{n+1}, y_i);
end
Set p_n(y_{n+1}) := 1/(n+1) Σ_{i=1}^{n+1} 1{t_i(y_i) ≥ t_{n+1}(y_{n+1})};
return 1{p_n(y_{n+1}) > α};
```

Theorem 4 (Vovk et al., 2005) *If the random variables $Y_1, \ldots, Y_{n+1} \sim Y$ are exchangeable, then a CP set is [type 1] valid, as in Definition 1.*

Proof This result is established in Vovk et al. (2005), but I provide an alternative explicit proof in the Appendix. The proof follows by first observing that a type 1 error is the event $\{p_n(y_{n+1}) \leq \alpha\}$, and then showing that $P\{p_n(Y_{n+1}) \leq \alpha\} \leq \alpha$. ■

While provably valid, the CP approach lacks the versatility to assign confidence to assertions $\{y_{n+1} : B \ni y_{n+1}\}$ if B does not coincide with a CP set at some level. In Section 3.3, a model-free formulation of GF inference is constructed that is able to assign GF-based probability the same as the level of a CP set (and is thus valid), but is also able to assign imprecise probabilities (i.e., lower and upper probabilities) to all other assertions. The necessary requisites on GF inference are introduced in the next section.

Definition 5 *The function $p_n(y) := \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{t_i(y_i) \geq t_{n+1}(y)\}$, as in Algorithm 3, will be referred to as the 'conformal p-value function'.*

3.2 GF inference

The motivating assumption for GF inference is an explicit association between data Y and an auxiliary variable U through some deterministic function G that depends on unknown population parameters of interest, θ. Expressed as,

$$Y = G(U, \theta), \tag{1}$$

the association is typically referred to as a *data generating equation*. A key aspect of the assumption is that the auxiliary variable has a completely known and fully specified distribution. The auxiliary variable can be understood similarly to the notion of a pivotal quantity that might be constructed in the context of statistical testing or bootstrapping. The goal is to build epistemic inference on the unknown θ by using the assumption of association (1).

From the GF inference perspective, the association (1) represents a mapping from a parameter space Θ to the support $\mathbb{Y}$ of the datum Y, and as such, once data are observed, an inverse mapping would contain valuable information about the unknown value θ. More

precisely, given an observed data set $y_1, \ldots, y_n$ generated independently from (1), there necessarily exists a corresponding set of auxiliary variable values $u_1, \ldots, u_n$ such that the unknown value θ solves the system of equations, $y_1 = G(u_1, \theta), \cdots, y_n = G(u_n, \theta)$. If the set of auxiliary values $u_1, \ldots, u_n$ were known, then this would be a deterministic inverse problem. Nonetheless, although $u_1, \ldots, u_n$ are unknown, it is assumed that the set comprises values that were generated independently and identically from the assumed known and fully specified distribution of the auxiliary variable U.

Accordingly, these facts motivate the formal definition of a GF distribution of θ, presented next in Definition 6. This definition is interpreted as: the probability that θ is contained in a set B is equal to the probability, under the distribution of $U_1, \ldots, U_n$, that the limiting quantity is equal to a θ that lies in the set B. Note that in this definition, $y_1, \ldots, y_n$ are regarded as fixed while $U_1, \ldots, U_n$ are random. Thus, the GF distribution is a distributional statistic for the unknown value θ, inheriting its uncertainty from the distribution of the auxiliary random variable, the same as $Y_1, \ldots, Y_n$. In Hannig et al. (2016) this is referred to as the *switching principle*. The notion of a distributional statistic for a fixed but unknown parameter, i.e., θ, is analogous to the role played by the posterior distribution in the Bayesian framework.

Definition 6 (Hannig et al., 2016) *Given an observed data set $y_1, \ldots, y_n$ generated independently from (1), a GF distribution on a parameter space Θ is defined as the weak limit,*

$$\lim_{\epsilon \to 0} \left\{ \operatorname*{argmin}_{\vartheta \in \Theta} \sum_{i=1}^{n} \|y_i - G(U_i, \vartheta)\|^2 \;\middle|\; \min_{\vartheta \in \Theta} \sum_{i=1}^{n} \|y_i - G(U_i, \vartheta)\|^2 \leq \epsilon \right\},$$

where G is a deterministic function, and the distribution of $U_1, \ldots, U_n$ is fully known and specified.

For discrete-valued data, the limit $\epsilon \to 0$ in Definition 6 reduces to setting $\epsilon = 0$, leading to an imprecise probability distribution over Θ. For example, in the case of binomial(m, θ) data, the data generating equation (1) may take the form:

$$Y = \sum_{k=1}^{m} 1\{U_k < \theta\},$$

where $U_1, \ldots, U_m \overset{\text{iid}}{\sim} \text{uniform}(0, 1)$. For an observed instance y from this data generating equation, the GF distribution for θ is obtained by replacing the unobserved $u_1, \ldots, u_m$ that generated y with an independent copy of the auxiliary variables $U_1^\star, \ldots, U_m^\star \overset{\text{iid}}{\sim} \text{uniform}(0, 1)$, and setting $\epsilon = 0$ in Definition 6. This leads to the imprecise GF distribution for θ defined by the interval-valued random variable of the form $(U_{(y)}^\star, U_{(y+1)}^\star] \subseteq \Theta$, where $U_1^\star, \ldots, U_m^\star \overset{\text{iid}}{\sim} \text{uniform}(0, 1)$ and $U_{(k)}^\star$ denotes the k-th order statistic of $U_1^\star, \ldots, U_m^\star$; i.e., $U_{(y)}^\star$ is a random variable characterizing the uncertainty about $u_{(y)}$—the unobserved auxiliary value corresponding to the observed data value y.

3.3 Model-free GF inference

The GF inference approach begins with the assumption that data are generated independently from a data generating equation as in (1). Such an assumption is model-based, and

requires explicit knowledge of the deterministic function G in (1) along with the distribution of the auxiliary variables. Instead, consider Assumption 7.

Assumption 7 *The variables $Y_1, \ldots, Y_{n+1} \in \mathbb{Y}$ are exchangeable and continuous (in the sense of having a continuous distribution function with respect to the Lebesgue measure).*

If an injective nonconformity measure Ψ can be constructed for the data, then under Assumption 7 a model-free data generating *association* for Y_{n+1} is given by:

$$\text{rank}\{t_{n+1}(Y_{n+1})\} = V \sim \text{uniform}\{1, \ldots, n+1\}, \tag{2}$$

where $\text{rank}\{t_{n+1}(Y_{n+1})\}$ denotes the position or *rank* of $t_{n+1}(Y_{n+1})$ in the order statistics (in ascending order) of the sample $t_1(Y_1), \ldots, t_{n+1}(Y_{n+1})$; i.e.,

$$\text{rank}\{t_j(y_j)\} := 1 + \sum_{i=1}^{n+1} 1\{t_j(y_j) > t_i(y_i)\},$$

for $j \in \{1, \ldots, n+1\}$. In this model-free approach, the phrase data generating *association* is used in place of data generating *equation* because knowledge of the true auxiliary variable value in equation (2) does *not* fully determine the datum value y_{n+1}. Nonetheless, the GF inference algorithm can be applied with reference to the datum variable $\text{rank}\{t_{n+1}(Y_{n+1})\}$, as usual, but for inference on y_{n+1}. First, replace the unobserved true auxiliary variable in (2) with an independent copy, $V^\star \sim \text{uniform}\{1, \ldots, n+1\}$. Second, apply the switching principle to obtain an imprecise GF distribution of the to-be-predicted value y_{n+1} as a distribution over the random sets,

$$A_n(V^\star) := \underset{y \in \mathbb{Y}}{\text{argmin}}\big\{|\text{rank}\{t_{n+1}(y)\} - V^\star|\big\} = \big\{y : \text{rank}\{t_{n+1}(y)\} = V^\star\big\}, \tag{3}$$

as illustrated in Figure 1.

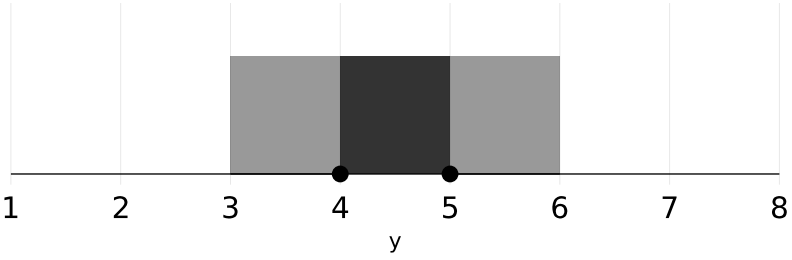

Figure 1: Hypothetical observed univariate data with $y_1 = 4$, $y_2 = 5$, and $n = 2$. With nonconformity measure $\Psi(y_{-i}^{n+1}, y_i) := |\text{mean}(y_{-i}^{n+1}) - y_i|$, the inner black region represents the values of y_{n+1} that would have rank 1 (i.e., $A_n(1) = \big\{y : \text{rank}[t_{n+1}(y)] = 1\big\}$), the outer grey region represents the values of $y_{n+1} \in A_n(2) = \big\{y : \text{rank}[t_{n+1}(y)] = 2\big\}$, and the outermost white region represents the values of $y_{n+1} \in A_n(n+1) = \big\{y : \text{rank}[t_{n+1}(y)] = n+1\big\}$.

Remark 8 *The continuity requirement of Assumption 7 together with the injectivity assumption about Ψ is to ensure that $t_i(Y_i) \neq t_j(Y_j)$ a.s. for any $i \neq j$. Otherwise, equation (2) is misspecified because the support of the random variable rank$\{t_{n+1}(Y_{n+1})\}$ may not include the entire set $\{1, \ldots, n+1\}$.*

Definition 9 (Shafer, 1976) *In DS theory, $\mu : 2^{\mathbb{Y}} \to [0,1]$ is called a basic probability assignment function if it satisfies $\mu(\emptyset) = 0$ and $\sum_{B \in 2^{\mathbb{Y}}} \mu(B) = 1$. With respect to the basic probability assignment function μ, a set $B \subseteq \mathbb{Y}$ is called a focal set if $\mu(B) > 0$.*

The imprecise GF probability measure (called the basic probability assignment function in DS theory) denoted by $\mu : 2^{\mathbb{Y}} \to [0,1]$ has focal sets

$$\mu\{A_n(V^\star)\} = \pi_v^n\left(V^\star = 1 + \sum_{i=1}^{n+1} 1\{t_{n+1}(y_{n+1}) > t_i(y_i)\}\right) = \frac{1}{n+1}, \tag{4}$$

where π_v^n denotes the discrete uniform probability mass function associated with the auxiliary variable $V^\star$.

An important insight from Figure 1 is that the imprecise GF distribution of y_{n+1} assigns, in particular, $1/(n+1)$ probability to the outermost region (beyond where any data were observed), and as such, it assigns $n/(n+1)$ probability to the complementary set (within which all of the data were observed). Moreover, a $k/(n+1)$ GF prediction set is given by,

$$\Omega_n(k) := \bigcup_{1 \le v \le k} A_n(v) = \big\{y \,:\, \text{rank}[t_{n+1}(y)] \le k\big\}, \tag{5}$$

for any $k \in \{1, \ldots, n+1\}$. Theorem 10, below, relates this model-free GF prediction set to a CP set via the *GF contour function*:

$$f_n(y) := \mu\big\{\Omega_n(V^\star) \ni y\big\}. \tag{6}$$

As a simple example, for the hypothetical data displayed in Figure 1, the GF contour is

$$f_n(y) = \begin{cases} 1 & \text{if } \ y \in (4,5) \\ \frac{2}{3} & \text{if } \ y \in (3,4] \cup [5,6) \\ \frac{1}{3} & \text{else} \end{cases}.$$

More interesting examples of GF contours, for synthetic Gaussian and Cauchy data, are plotted in Figure 2. The important implication of Theorem 10 is that $\Upsilon_n^\alpha := \{y : f_n(y) > \alpha\}$ is a [type 1] valid, model-free GF prediction set, as stated in Corollary 11.

At any level $\alpha \in (0,1)$, the region Υ_n^α is easily determined in the plots in Figure 2 by drawing a horizontal line at the value of α and including all values of y that satisfy $f_n(y) > \alpha$ (i.e., Υ_n^α is a pre-image set of f_n). Although a contour is *not* understood as a density function, construction of Υ_n^α is akin to the construction of highest posterior density credible sets in Bayesian inference.

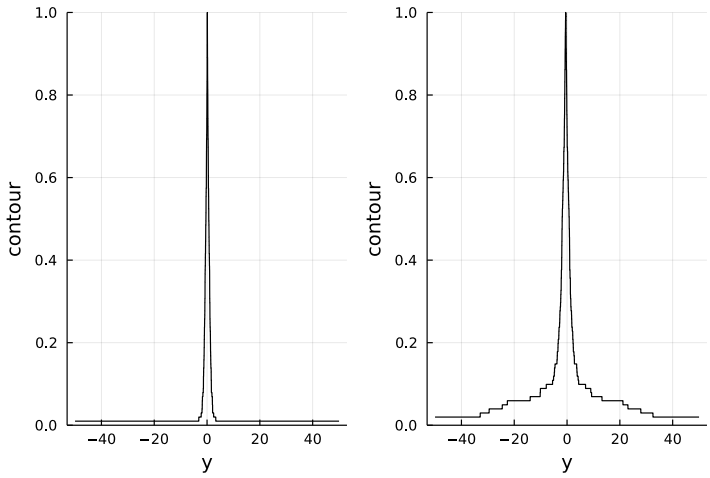

Figure 2: Both panels display plots of the GF contour, $f_n(y) = \mu\{\Omega_n(V^\star) \ni y\}$; the left and right plots are based on samples of $n = 100$ realizations from the standard Gaussian and standard Cauchy distributions, respectively.

Theorem 10 *Under Assumption 7 the GF contour $f_n(y)$ is a conformal p-value function.*

Proof This result is simply a statement of the fact that using $f_n(y_{n+1})$ in place of $p_n(y_{n+1})$ in Algorithm 3 defines a conformal algorithm. This follows because

$$\begin{aligned} f_n(y_{n+1}) &= \mu\big\{\Omega_n(V^\star) \ni y_{n+1}\big\} \\ &= \mu\big\{\text{rank}\{t_{n+1}(y_{n+1})\} \leq V^\star\big\} \\ &= \frac{n+1-\text{rank}\{t_{n+1}(y_{n+1})\}+1}{n+1} \\ &= \frac{n+1-1-\sum_{i=1}^{n+1} 1\{t_{n+1}(y_{n+1}) > t_i(y_i)\}+1}{n+1} \\ &= \frac{\sum_{i=1}^{n+1} 1\{t_{n+1}(y_{n+1}) \leq t_i(y_i)\}}{n+1}. \end{aligned} \tag{7}$$

■

Corollary 11 *Under Assumption 7 the GF prediction set Υ_n^α is [type 1] valid, as in Definition 1.*

Proof As shown in Theorem 10, $f_n(y_{n+1})$ is equivalent to $p_n(y_{n+1})$ in Algorithm 3, and so

$$P(\Upsilon_n^\alpha \not\ni Y_{n+1}) = P\big(f_n(Y_{n+1}) \leq \alpha\big) \leq \alpha,$$

as a direct consequence of Theorem 4. ■

It is clear from Theorem 10 that CP sets can be constructed from the GF inference paradigm, namely Υ_n^α is a CP set. Furthermore, it follows from expression (7) that any

CP set as constructed from Algorithm 3 can be understood as a union of focal sets from the imprecise GF probability distribution of y_{n+1} defined by (3). This fact establishes the fundamental connection between GF inference and CP.

4. Characterizing the imprecision in the model-free GF distribution

The imprecision in the GF distribution is that the probability mass μ is only defined for sets of values $A_n(v) \subseteq \mathbb{Y}$, for $v \in \{1, \ldots, n+1\}$, rather than a mass or density function defined for all points in $\mathbb{Y}$, as in the precise probability scenario; and the imprecision comes from the fact that nothing is being assumed about the underlying distribution of the data. The novelty of the approach is that the GF framework is, nonetheless, versatile enough to provide inferences and construct CP sets, based on the model-free association (2) with the sole assumption of exchangeable, continuous data. Inferences can be facilitated by the construction of belief and plausibility functions, denoted $\underline{\mu}$ and $\overline{\mu}$, respectively, so that for any event $B \subseteq \mathbb{Y}$ pertaining to the prediction of y_{n+1}, i.e., $\{y_{n+1} : B \ni y_{n+1}\}$,

$$\overline{\mu}(B) := \sum_{j=1}^{n+1} \mu\{A_n(j)\} \cdot 1\big\{A_n(j) \cap B \neq \emptyset\big\}, \text{ and}$$
$$\underline{\mu}(B) := \sum_{j=1}^{n+1} \mu\{A_n(j)\} \cdot 1\big\{A_n(j) \subseteq B\big\}.$$

A probability measure Δ is naturally considered *compatible* with μ if for every measurable set B, $\underline{\mu}(B) \leq \Delta(B) \leq \overline{\mu}(B)$. The set of all such probability measures is called the *credal set* of μ, and can be expressed as

$$\mathscr{C}(\mu) := \big\{\Delta : \Delta(B) \leq \overline{\mu}(B), \text{ for any measurable set } B\big\}.$$

Further, this construction implies that any probability measure $\Delta \in \mathscr{C}(\mu)$ must assign the same probability mass as μ to any focal set of μ. This fact is established as a direct consequence of Lemma 12.

Lemma 12 *A probability measure $\Delta \in \mathscr{C}(\mu)$ if and only if $\Delta\{A_n(v)\} = \frac{1}{n+1} = \mu\{A_n(v)\}$, for every $v \in \{1, \ldots, n+1\}$.*

Proof First, suppose that $\Delta \in \mathscr{C}(\mu)$. Then for any $v \in \{1, \ldots, n+1\}$, using the fact that $A_n(1), \ldots, A_n(n+1)$ are mutually disjoint,

$$\begin{aligned}
\mu\{A_n(v)\} &= \sum_{j=1}^{n+1} \mu\{A_n(j)\} 1\big\{A_n(j) \subseteq A_n(v)\big\} \\
&= \underline{\mu}\{A_n(v)\} \\
&\leq \Delta\{A_n(v)\} \\
&\leq \overline{\mu}\{A_n(v)\} \\
&= \sum_{j=1}^{n+1} \mu\{A_n(j)\} 1\big\{A_n(j) \cap A_n(v) \neq \emptyset\big\} \\
&= \mu\{A_n(v)\}
\end{aligned}$$

The desired result follows by equation (4).

For the converse direction, assume that $\Delta\{A_n(v)\} = \frac{1}{n+1}$, for every $v \in \{1, \ldots, n+1\}$. Then, for any measurable set B, using the fact that $A_n(1), \ldots, A_n(n+1)$ are mutually disjoint and collectively exhaustive over $\mathbb{Y}$,

$$\begin{aligned}
\Delta(B) &= \sum_{v=1}^{n+1} \Delta\big\{B \cap A_n(v)\big\} \\
&\leq \sum_{v=1}^{n+1} \Delta\{A_n(v)\} 1\big\{A_n(v) \cap B \neq \emptyset\big\} \\
&= \sum_{v=1}^{n+1} \frac{1}{n+1} 1\big\{A_n(v) \cap B \neq \emptyset\big\} \\
&= \sum_{v=1}^{n+1} \mu\{A_n(v)\} 1\big\{A_n(v) \cap B \neq \emptyset\big\} \\
&= \overline{\mu}(B).
\end{aligned}$$

■

The implication of interest of Lemma 12 is that any probability measure compatible with the imprecise GF probability measure μ must assign uniform (i.e., $\frac{1}{n+1}$) probability to each of the mutually disjoint and collectively exhaustive regions $A_n(1), \ldots, A_n(n+1)$. Assuming a density function exists, the probability density associated with any $\Delta \in \mathscr{C}(\mu)$, however, can have arbitrary shape over each region $A_n(v)$, subject to the constraint that it integrates to $\frac{1}{n+1}$.

4.1 Connection to nonparametric predictive inference

A special case of the model-free GF inference construction is the NPI construction developed by Coolen and his collaborators since the late 1990s (e.g., Coolen, 1998; Augustin and Coolen, 2004; Coolen and Coolen-Maturi, 2024), and the so-called "Dempster-Hill procedure" based on the seminal papers Dempster (1963); Hill (1968, 1988). The name *Dempster-Hill procedure* is coined in Vovk (2024) and is described as the statement (paraphrased from Vovk, 2024):

> Given data $y_1, \ldots, y_n$, the probability that the next observation y falls in $(y_{(i)}, y_{(i+1)})$ is $1/(n+1)$, for each $i \in \{0, \ldots, n\}$; by definition, $y_{(0)} := -\infty$ and $y_{(n+1)} := \infty$.

This idea, taken to be the fundamental assumption of NPI, is argued in Vovk (2024) to trace back as far as Jeffreys (1932). Taking the nonconformity score to be each datum itself, i.e., $t(Y_i) := Y_i$, and deriving the model-free GF distribution exactly yields the Dempster-Hill procedure, by virtue of equation (4). From here, all of NPI follows via the same imprecise probability calculus of lower and upper probabilities described above (Coolen and Coolen-Maturi, 2024), but is generalized by model-free GF for arbitrary choices of nonconformity score.

The Dempster-Hill procedure is used in Vovk (2024) to construct the "Dempster-Hill pivot", a continuity-corrected conformal p-value function. This pivot is regarded by Vovk as an approximate probability distribution function and is used to define a "conformal predictive distribution". Such a conformal predictive distribution is defined for nonconformity scores such that the conformal p-value function is an increasing function of the to-be-predicted y, i.e., also generalizing the Dempster-Hill procedure. The model-free GF paradigm does not, however, rely on monotonicity of the nonconformity score because it avoids direct construction of a probability distribution function. This is avoided by explicitly dealing with the imprecise probabilities that arise from the fiducial switching arguments, as discussed from the beginning of Section 4. GF and conformal predictive distributions can be regarded as two philosophically different but related interpretations of fiducial predictions (Vovk embraced such a perspective in Vovk, 2024).

In the context of the Dempster-Hill procedure, the next two results demonstrate the non-asymptotic, sub-exponential concentration that establishes the root-n consistency of all probability measures $\Delta \in \mathscr{C}(\mu)$. In particular, Theorem 13 demonstrates that pointwise, $|\Delta\{(-\infty, y]\} - F(y)| = o_p(n^{-\gamma})$, for any $\gamma \in [0, .5)$, where F is the distribution function associated with the true distribution of $Y_{n+1} \sim P$. Theorem 14 establishes an even faster rate of convergence on the focal sets; $|\Delta\{A_n(v)\} - P\{A_n(v)\}| = o_p(n^{-\tau})$, for any $\tau \in [0, 1)$. These results are expected from well-established consistency properties of the empirical distribution function, in the context of empirical process theory. Namely, the disjoint focal sets $A_n(1), \ldots, A_n(n+1)$ will be narrower and clustered in regions of high probability density, and wider and fewer in regions of low probability density.

Theorem 13 *Let $Y_1, \ldots, Y_n \overset{iid}{\sim} P$ be a collection of continuous random variables and $t_i(Y_i) := Y_i$ for $i \in \{1, \ldots, n\}$. For any $\Delta \in \mathscr{C}(\mu)$, $y \in \mathbb{R}$, $\gamma \in [0, .5)$, $\epsilon > 0$, and $n > 4n^{\gamma}/\epsilon - 1$,*

$$P\Big(n^{\gamma}\big|\Delta\{(-\infty, y]\} - F(y)\big| > \epsilon\Big) \leq 4e^{-\frac{\epsilon^2}{8}n^{1-2\gamma}} \cdot 1\{F(y) > 0\}.$$

Proof Denote $B := (-\infty, y]$. For any $\Delta \in \mathscr{C}(\mu)$, by definition, $\underline{\mu}(B) \leq \Delta(B) \leq \overline{\mu}(B)$, and so

$$|\Delta(B) - P(B)| \leq \max\big\{|\underline{\mu}(B) - P(B)|, |\overline{\mu}(B) - P(B)|\big\}.$$

A union bound then gives

$$P\Big(n^{\gamma}\big|\Delta\{(-\infty, y]\} - F(y)\big| > \epsilon\Big) \leq P\Big(n^{\gamma}\big|\underline{\mu}(B) - P(B)\big| > \epsilon\Big) + P\Big(n^{\gamma}\big|\overline{\mu}(B) - P(B)\big| > \epsilon\Big),$$

which suggests that the desired result follows by establishing the stochastic concentration of the belief and plausibility functions around $P(B) = F(y)$. The latter concentration is established next.

Let M_n be the number of focal sets having a nonempty intersection with B:

$$M_n := \big|\big\{v \,:\, (-\infty, y] \cap A_n(v) \neq \emptyset\big\}\big|.$$

Due to the choice of $t_i(Y_i) := Y_i$ for the nonconformity score, $A_n(1) = (-\infty, Y_{(1)}]$, $A_n(2) = (Y_{(1)}, Y_{(2)}]$, $\ldots$, $A_n(n+1) = (Y_{(n)}, \infty)$, and so $M_n = 1 + \sum_{i=1}^{n} 1\{Y_i \leq y\}$, which implies

that $Z_n := M_n - 1 \sim \text{binomial}(n, p_B)$, where $p_B := P(B) = F(y)$. Then, by definition, $\overline{\mu}(B) = M_n/(n+1)$. Accordingly,

$$\begin{aligned} P\Big(n^\gamma\big|\overline{\mu}(B) - P(B)\big| > \epsilon\Big) &= P\left(\left|\frac{M_n}{n+1} - p_B\right| > \frac{\epsilon}{n^\gamma}\right) \\ &\leq 1\left\{\frac{1}{n+1} > \frac{\epsilon}{2n^\gamma}\right\} + P\left(\left|\frac{M_n - 1}{n+1} - p_B\right| > \frac{\epsilon}{2n^\gamma}\right) \\ &= P\left(\left|\frac{Z_n}{n+1} - p_B\right| > \frac{\epsilon}{2n^\gamma}\right), \end{aligned} \tag{8}$$

where the final equivalence follows by assumption. If $p_B = F(y) = 0$, then all of the above terms are 0 due to the fact that $Z_n = 0$ with probability $(1 - p_B)^n = 1$. Conversely, if $p_B = F(y) > 0$, then by the triangle inequality

$$\begin{aligned} P\Big(n^\gamma\big|\overline{\mu}(B) - P(B)\big| > \epsilon\Big) &\leq P\left(|Z_n - n \cdot p_B| + p_B > \frac{(n+1)\epsilon}{2n^\gamma}\right) \\ &\leq P\left(|Z_n - E(Z_n)| > \frac{(n+1)\epsilon}{4n^\gamma}\right) + 1\left\{p_B > \frac{(n+1)\epsilon}{4n^\gamma}\right\} \\ &\leq 2e^{-2\{\frac{(n+1)\epsilon}{4n^\gamma}\}^2/n} + 0, \end{aligned}$$

where the last approximation is an application of the Hoeffding inequality (regarding Z_n as the sum of n independent Bernoulli(p_B) random variables), and the assumption that $n > 4n^\gamma/\epsilon - 1$. The more concise expression in the theorem statement employs the fact that $n + 1 \geq n \geq 1$ implies $(n+1)(\frac{n+1}{n}) \geq n + 1 \geq n$, so that

$$e^{-2\{\frac{(n+1)\epsilon}{4n^\gamma}\}^2/n} = e^{-2\frac{\epsilon^2}{16n^{2\gamma}}(n+1)(\frac{n+1}{n})} \leq e^{-\frac{\epsilon^2}{8n^{2\gamma}}n} = e^{-\frac{\epsilon^2}{8}n^{1-2\gamma}}.$$

Next, employ a similar argument to establish the concentration of the belief function:

$$\big|\{v : A_n(v) \subseteq (-\infty, y]\}\big| = \sum_{i=1}^{n} 1\{Y_i \leq y\} = Z_n \sim \text{binomial}(n, p_B),$$

so that, by definition, $\underline{\mu}(B) = Z_n/(n+1)$. Then

$$P\Big(n^\gamma\big|\underline{\mu}(B) - P(B)\big| > \epsilon\Big) = P\left(\left|\frac{Z_n}{n+1} - p_B\right| > \frac{\epsilon}{n^\gamma}\right) \leq 2e^{-\frac{\epsilon^2}{2}n^{1-2\gamma}},$$

where the inequality is established using the same arguments following after the final expression in equation (8). Using the fact that $1/2 > 1/8$, it follows that both the belief and plausibility stochastic concentrations can be bounded above by $2e^{-\frac{\epsilon^2}{8}n^{1-2\gamma}}$. Adding these two bounds, as indicated by the union bound derived at the beginning of the proof, establishes the theorem statement. ■

Theorem 14 *Let $Y_1, \ldots, Y_n \overset{iid}{\sim} P$ be a collection of continuous random variables and $t_i(Y_i) := Y_i$ for $i \in \{1, \ldots, n\}$. For any $\Delta \in \mathscr{C}(\mu)$, $v \in \{1, \ldots, n+1\}$, $\tau \in [0, 1)$, and $\epsilon > 0$,*

$$P\Big(n^\tau \big|\Delta\{A_n(v)\} - P\{A_n(v)\}\big| > \epsilon\Big) = 1 - (1 - b_n)^n + (1 - c_n)^n,$$

where $b_n := \max\big\{\frac{1}{n+1} - \frac{\epsilon}{n^\tau}, 0\big\}$, $c_n := \min\big\{\frac{1}{n+1} + \frac{\epsilon}{n^\tau}, 1\big\}$. In particular,

$$1 - (1 - b_n)^n + (1 - c_n)^n \leq e^{-n^{1-\tau}\epsilon},$$

for all $n > \max\big\{\frac{n^\tau - \epsilon}{\epsilon}, \frac{\epsilon}{n^\tau - \epsilon}\big\}$.

Proof By continuity and independence, denoting $Y_{(0)} := -\infty$ and $Y_{(n+1)} := \infty$, for any $v \in \{1, \ldots, n+1\}$, $W_n := F(Y_{(v)}) - F(Y_{(v-1)}) \sim \text{beta}(1, n)$. Then, by Lemma 12,

$$\begin{aligned} P\Big(n^\tau \big|\Delta\{A_n(v)\} - P\{A_n(v)\}\big| > \epsilon\Big) &= P\left(n^\tau \left|\frac{1}{n+1} - \big[F(Y_{(v)}) - F(Y_{(v-1)})\big]\right| > \epsilon\right) \\ &= 1 - P\left(\left|W_n - \frac{1}{n+1}\right| \leq \frac{\epsilon}{n^\tau}\right) \\ &= 1 - P\big(b_n \leq W_n \leq c_n\big). \end{aligned} \tag{9}$$

Next,

$$P\big(b_n \leq W_n \leq c_n\big) = \int_{b_n}^{c_n} n(1-x)^{n-1}\,dx = -(1-x)^n \Big|_{b_n}^{c_n} = (1 - b_n)^n - (1 - c_n)^n.$$

Finally, for all $n > \max\big\{\frac{n^\tau - \epsilon}{\epsilon}, \frac{\epsilon}{n^\tau - \epsilon}\big\}$, it follows that $b_n = 0$ and

$$1 - (1 - b_n)^n + (1 - c_n)^n = (1 - c_n)^n \leq e^{-nc_n} \leq e^{-n^{1-\tau}\epsilon},$$

where the first inequality follows by the property of the exponential function that $1 + x \leq e^x$ for all $x \in \mathbb{R}$. In particular, set $x = -c_n$ and then, noting that $c_n \in (0, 1)$, take the n-th power of both sides. The second inequality follows from $nc_n = \frac{n}{n+1} + \frac{n\epsilon}{n^\tau} \geq n^{1-\tau}\epsilon$. ■

4.2 Connection to the inferential models framework

A related construction to the model-free GF setup is the *nonparametric IM* construction introduced in Cella and Martin (2022). The difference is that, whereas the model-free GF is determined by a basic probability assignment function over the *disjoint* random focal sets $A_n(V^\star)$, the nonparametric IM is determined via a contour function over *nested* predictive random sets. In the notation of Section 3.3 the nonparametric IM is defined by the contour function $f_n(y) = \mu\big\{\Omega_n(V^\star) \ni y\big\}$, recalling $\Omega_n(\cdot)$ as defined in (5), which is the same as the GF contour function given in (6). Furthermore, whereas the GF plausibility and belief functions are given by $\overline{\mu}(\cdot)$ and $\underline{\mu}(\cdot)$, respectively, in Section 4, the IM plausibility and belief

functions can be expressed as:

$$\overline{\zeta}(B) := \sum_{j=1}^{n+1} \mu\{A_n(j)\} \cdot 1\big\{\Omega_n(j) \cap B \neq \emptyset\big\}, \text{ and}$$
$$\underline{\zeta}(B) := \sum_{j=1}^{n+1} \mu\{A_n(j)\} \cdot 1\big\{\Omega_n(j) \subseteq B\big\}.$$

The nested random sets in the IM construction result from the IM being built to satisfy a notion of validity for imprecise probabilities, as stated next in Definition 15.

Definition 15 (Type 2 validity – Cella and Martin, 2022) *Let $\overline{\Pi}$ be an upper probability constructed from observed data $y_1, \ldots, y_n \sim Y$, and assume $y_{n+1} \sim Y$. Denoting by P the probability measure associated with Y, the lower and upper probability pair $(\underline{\Pi}, \overline{\Pi})$ is type 2 valid if, for all (α, n, P), $P\big\{\overline{\Pi}(B) \leq \alpha$ and $Y_{n+1} \in B$ for some $B\big\} \leq \alpha$.*

The utility of an imprecise probability that satisfies type 2 validity is a guarantee that for any event B containing the true value of the quantity to-be-predicted (e.g., corresponding to a null hypothesis), the upper probability (be it a measure of plausibility, possibility, etc.) assigned to B will not be too small too often (over replications of $y_1, \ldots, y_n, y_{n+1}$). This ensures finite-sample type 1 error protection for general purpose imprecise probabilistic inference, i.e., without pre-specifying or fixing an event/hypothesis B.

Mathematically, type 2 validity is achieved by the IM because $\overline{\zeta}(\cdot)$ is a *consonant* plausibility function. A consonant plausibility function is a plausibility function $\overline{\Pi}$ that satisfies $\overline{\Pi}(B) = \sup_{y \in B} f(y)$, for any $B \subseteq \mathbb{Y}$, for some function $f : \mathbb{Y} \to [0, 1]$ with $\sup_{y \in \mathbb{Y}} f(y) = 1$. It is established on page 122 of Cella and Martin (2022) that $\overline{\zeta}(B) = \sup_{y \in B} f_n(y)$, where again, $f_n(\cdot)$ also appears as the GF contour function in (6) used to construct the CP sets from the model-free GF distribution. Therein lies the GF–IM–CP connection.

Notice that $(\underline{\zeta}, \overline{\zeta})$ are constructed here from the GF mass function $\mu(\cdot)$, demonstrating that a consonant plausibility function—and thus a type 2 valid imprecise probabilistic predictor—is available within the model-free GF context, as well (consonance was also implicit in establishing the GF-CP connection in Theorem 10). Note that it is the belief function $\underline{\zeta}(B) = 1 - \overline{\zeta}(B^c) = 1 - \sup_{y \in B^c} f_n(y)$ that yields $k/(n+1)$ when B is a $k/(n+1)$ level CP set, i.e., when $B = \Omega_n(k)$.

Type 2 validity, however, is not satisfied by $(\underline{\mu}, \overline{\mu})$ based on the following argument. The focal sets $A_n(1), \ldots, A_n(n+1)$ are mutually disjoint and collectively exhaustive, and so for any $y \in \mathbb{Y}$, there exists a single $j \in \{1, \ldots, n+1\}$ such that $y \in A_n(j)$. Accordingly, $\overline{\mu}(\{y\}) = \overline{\mu}\{A_n(j)\} = 1/(n+1)$, which implies that for any $\alpha \in [1/(n+1), 1)$,

$$\begin{aligned} \alpha &< 1 \\ &= P\big[\overline{\mu}(\{Y_{n+1}\}) \leq \alpha \text{ and } Y_{n+1} \in \{Y_{n+1}\}\big] \\ &\leq P\big\{\overline{\mu}(B) \leq \alpha \text{ and } Y_{n+1} \in B \text{ for some } B\big\}. \end{aligned}$$

Thus, while type 2 validity is satisfied for belief and plausibility pairs constructed from $\Omega_n(V^\star)$, i.e., the union of $A_n(1), \ldots, A_n(V^\star)$, it is not satisfied for belief and plausibility pairs constructed directly from $A_n(V^\star)$. This demonstration of type 2 *in*validity of the

model-free GF construction for general sets is an instance of the philosophy set forth by 'safe probability' of Grünwald (2018): one can use the GF method for making valid probabilistic assessments of certain events, but applying it to arbitrary events may not make sense.

There do, nonetheless, remain utilities to imprecise inference based on $(\underline{\mu}, \overline{\mu})$. First, even though it is not type 2 valid, it does guarantee type 1 error protection for events B that correspond to CP sets, as established by the developments in Section 3.3. Second, consistency, as in Theorems 13 and 14, follows more directly from the $(\underline{\mu}, \overline{\mu})$ construction versus the $(\underline{\zeta}, \overline{\zeta})$ construction: a clear advantage of model-free GF versus nonparametric IM. Third, based on the characterization of Lemma 12, the $(\underline{\mu}, \overline{\mu})$ construction makes it easy to find pignistic transformations, i.e., the credal set $\mathscr{C}(\mu)$ simply consists of all probability measures assigning $1/(n+1)$ probability over each $A_n(v)$. Such a pignistic transformation is illustrated, and others are discussed, in the following section. Furthermore, using this illustration, Figures 4 and 5 give a visual depiction of the differences between the (consonant) nonparametric IM $(\underline{\zeta}, \overline{\zeta})$ versus the model-free GF $(\underline{\mu}, \overline{\mu})$. Moreover, Section 5 is a case study illustration of the advantages of model-free GF for both deriving pignistic transformations and establishing statistical consistency, compared to nonparametric IM.

5. Illustration of a pignistic transformation

The preceding developments make a case for the essential role of imprecise probabilistic reasoning in finite-sample, valid inference. If, however, a decision maker is pressed to make a decision, a transformation from the credal level to the pignistic level might be considered, i.e., the choice of a single probability distribution from the credal set. Such transformations have been considered previously (e.g., see Smets, 1990; Dubois et al., 2004), but no general theory for their construction/evaluation exists (Grünwald, 2018). Rather than attempting to propose general principles here, this section serves to simply illustrate how such a transformation can be constructed based on the well-studied concept of a maximum entropy distribution (MED) with respect to the Lebesgue measure. The optimality of such a concept in this context is postponed to future work. Nonetheless, the usual intuition for maximizing entropy is a desire to be least informative or maximally ignorant with respect to unknown information.

It is well-known that the MED over a bounded interval is uniform, and so the MED over the credal set $\mathscr{C}(\mu)$ should have a density π_y^n that is flat over each focal set $A_n(v)$. Such a construction might require the modification of $A_n(1)$ and/or $A_n(n+1)$ so that the support of π_y^n is restricted to $[\kappa_n^{\min}, \kappa_n^{\max}]$ for arbitrarily small/large data-dependent choices of $\kappa_n^{\min}$ and $\kappa_n^{\max}$, so that uniform densities will integrate over these focal regions. The MED is derived by integrating the conditional uniform density on every focal set with respect to the GF measure associated with the auxiliary variable: for $y \in [\kappa_n^{\min}, \kappa_n^{\max}]$,

$$\pi_y^n(y) = \sum_{v=1}^{n+1} \pi_{y|v}^n(y \mid v) \cdot \pi_v^n(v) = \sum_{v=1}^{n+1} \begin{cases} \frac{1}{\lambda\{A_n(v)\}} \cdot \frac{1}{n+1} 1\{y \in A_n(v)\} & \text{if } |A_n(v)| > 1 \\ \delta_{A_n(v)}(y) \cdot \frac{1}{n+1} & \text{if } |A_n(v)| = 1 \end{cases}, \quad (10)$$

where λ is the Lebesgue measure on $\mathbb{Y}$, $|\cdot|$ denotes the cardinality of a set-valued argument, and $\delta_{A_n(v)}(\cdot)$ is the Dirac delta function for a singleton set $A_n(v)$. It is established shortly, by Lemma 17 and Theorem 18, that π_y^n is in fact the density associated with the MED over

$\mathscr{C}(\mu)$. Moreover, sampling from this distribution is intuitive, and is described by Algorithm 16; see Figure 3 for an empirical illustration.

Algorithm 16: Sampling according to the MED density π_y^n from equation (10).

Input: Prediction regions $A_n(1), \ldots, A_n(n+1)$.
Output: A realized instance of the random variable with density function π_y^n.
1 Sample $v^\star \sim \text{uniform}\{1, \ldots, n+1\}$;
2 Sample $y^\star \sim \text{uniform}\{A_n(v^\star)\}$;
3 **return** $y^\star$;

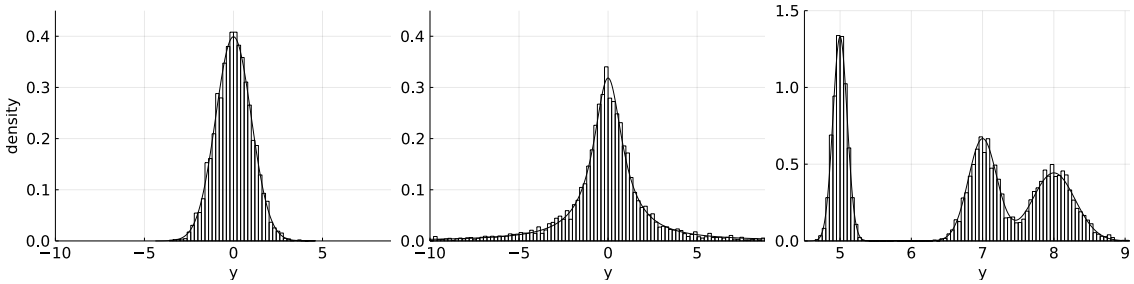

Figure 3: Histograms of samples from π_y^n computed by Algorithm 16, and based on data sets of size $n = 10,000$ from the standard Gaussian distribution (left panel), standard Cauchy distribution (middle panel), and a mixture distribution (right panel). The nonconformity measure is $t(y_i) := y_i$. The black lines are plots of the respective density functions associated with the data.

Lemma 17 *The probability measure* $\Pi_y^n(\cdot) := \int_{(\cdot)} \pi_y^n(y)\, dy \in \mathscr{C}(\mu)$.

Proof For any $j \in \{1, \ldots, n+1\}$,

$$\Pi_y^n\{A_n(j)\} = \int_{A_n(j)} \pi_y^n(y)\, dy = \begin{cases} \int_{A_n(j)} \frac{1}{\lambda\{A_n(j)\}} \cdot \frac{1}{n+1}\, dy = \frac{1}{n+1} & \text{if } |A_n(j)| > 1 \\ \int_{A_n(j)} \delta_{A_n(j)}(y) \cdot \frac{1}{n+1}\, dy = \frac{1}{n+1} & \text{if } |A_n(j)| = 1 \end{cases}.$$

Thus, $\Pi_y^n \in \mathscr{C}(\mu)$ as a consequence of Lemma 12. ■

Theorem 18 *The probability distribution having density function* π_y^n *is the MED over all probability measures in* $\mathscr{C}(\mu)$*, supported on* $[\kappa_n^{\min}, \kappa_n^{\max}]$.

Proof As illustrated by Lemma 12, the MED has a density residing in the set of density functions $\mathscr{Q}$ such that for every $q \in \mathscr{Q}$ and for every $v \in \{1, \ldots, n+1\}$,

$$\int_{A_n(v)} q(y)\, dy = \frac{1}{n+1}.$$

If $|A_n(v)| = 1$, then $q(y) = \delta_{A_n(v)}(y) \cdot \frac{1}{n+1}$ over $A_n(v)$. Alternatively, over the non-singleton focal set regions, the MED over $\mathscr{C}(\mu)$ can be found via the method of Lagrange multipliers

constrained to the set $\mathscr{Q}$. The constrained entropy functional has the form

$$J[q] = -\int_{\kappa_n^{\min}}^{\kappa_n^{\max}} q(y)\log\{q(y)\}\,dy + \sum_{j\,:\,|A_n(j)|>1} \beta_j \left[\int_{A_n(j)} q(y)\,dy - \frac{1}{n+1}\right],$$

and can be minimized using standard techniques from calculus of variations (a standard text on this subject is Gelfand and Fomin, 2000). The first-order condition for an optimum, based on the functional derivative is

$$\frac{\delta J}{\delta q} = -\log\{q(y)\} - 1 + \sum_{j\,:\,|A_n(j)|>1} \beta_j 1\{y \in A_n(j)\} = 0.$$

Thus, the MED density has the form $q(y) = e^{-1+\sum_{j\,:\,|A_n(j)|>1}\beta_j 1\{y\in A_n(j)\}}$, subject to the constraint

$$\frac{1}{n+1} = \int_{A_n(v)} e^{-1+\sum_{j\,:\,|A_n(j)|>1}\beta_j 1\{y\in A_n(j)\}}\,dy = \int_{A_n(v)} e^{-1+\beta_v}\,dy = e^{-1+\beta_v}\lambda\{A_n(v)\},$$

and so $q(y) = \frac{1}{\lambda\{A_n(v)\}}\cdot\frac{1}{n+1}$ for $y \in A_n(v)$ for every $v \in \{1,\ldots,n+1\}$. Therefore,

$$\begin{aligned} q(y) &= \sum_{v=1}^{n+1} \begin{cases} \frac{1}{\lambda\{A_n(v)\}}\cdot\frac{1}{n+1} 1\{y \in A_n(v)\} & \text{if } |A_n(v)|>1 \\ \delta_{A_n(v)}(y)\cdot\frac{1}{n+1} & \text{if } |A_n(v)|=1 \end{cases} \\ &= \pi_y^n(y). \end{aligned}$$

■

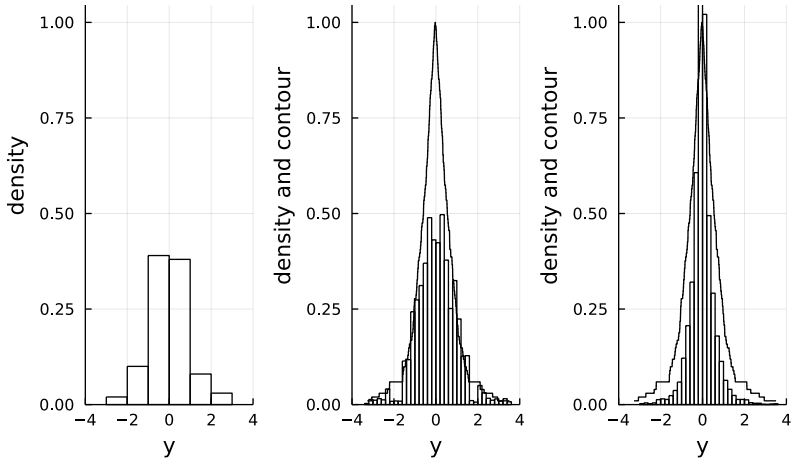

Figure 4: Based on the $n = 100$ data points drawn from the standard Gaussian distribution, as summarized by the histogram in the left panel, the middle and right panels display histograms of samples of size 10,000 drawn, respectively, from Algorithm 16 and its analogue by replacing $A_n(v^\star)$ (corresponding to model-free GF) with $\Omega_n(v^\star)$ (corresponding to nonparametric IM, i.e., consonant). The nonconformity measure is $t_i(y_i) := |\text{mean}(y_{-i}^{n+1}) - y_i|$. For reference, the contour function is provided as the black line in the middle and right panels.

An analogous sampling procedure to Algorithm 16 can be constructed by sampling from $\Omega_n(v^\star)$ rather than $A_n(v^\star)$ in Algorithm 16. A comparison of both such algorithms in Figures 4 and 5 gives a visual depiction of the differences between the consonant IM construction $(\underline{\zeta}, \overline{\zeta})$ versus $(\underline{\mu}, \overline{\mu})$. The point of these figures is to highlight the modal tendency of the consonance property.

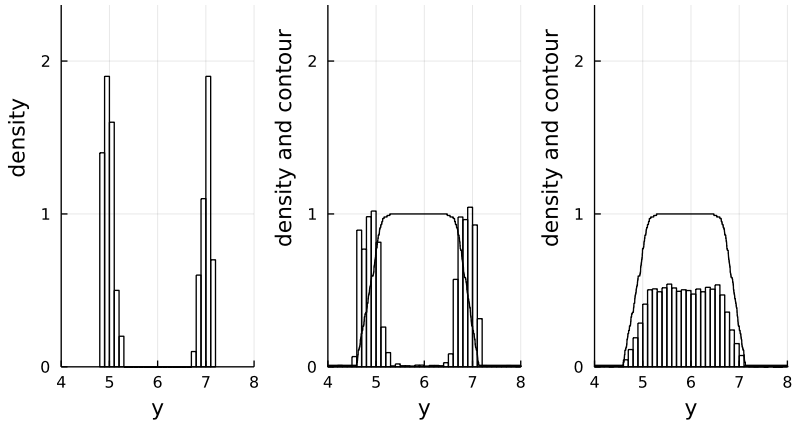

Figure 5: Based on the $n = 100$ data points drawn from a mixture of two Gaussian distributions, as summarized by the histogram in the left panel, the middle and right panels display histograms of samples of size 10,000 drawn, respectively, from Algorithm 16 and its analogue by replacing $A_n(v^\star)$ (corresponding to model-free GF) with $\Omega_n(v^\star)$ (corresponding to nonparametric IM, i.e., consonant). The nonconformity measure is $t_i(y_i) := |\text{mean}(y_{-i}^{n+1}) - y_i|$. For reference, the contour function is provided as the black line in the middle and right panels.

5.1 Other pignistic transformations from the GF literature

Although heuristic methods have been discussed and evaluated empirically (Hannig, 2009), mapping GF imprecise distributions to precise distributions has remained largely an open question for research in the GF literature. Recall the example of the imprecise GF distribution constructed for the parameter of a binomial distribution in Section 3.2. The existing GF inference literature suggests mapping the imprecise GF distribution to a precise distribution by taking some point in the interval $(U^\star_{(y)}, U^\star_{(y+1)}]$, which can be expressed as suggested in Hannig (2009) as $\mathcal{R}_\theta(y) = U^\star_{(y)} + D(U^\star_{(y+1)} - U^\star_{(y)})$, for some random variable D supported on or in $[0, 1]$. There are five options for the choice of D discussed in Hannig (2009). The first is the maximum entropy $D \sim \text{uniform}(0, 1)$, and the second is the maximum variance $D \sim \text{uniform}\{0, 1\}$, which amounts to arithmetic averaging of the densities of the endpoints. The third choice is $D \sim \text{beta}(0.5, 0.5)$, which leads to the Bayesian posterior of $\mathcal{R}_\theta(y)$ using the Jeffreys prior. This also corresponds to the geometric mean of the densities of the endpoints, and is advocated for by Schweder and Hjort (2016). The fourth choice is

$$D \mid U_1^\star, \ldots, U_n^\star = \begin{cases} 0 & \text{with probability } U^\star_{(y)} \\ 1 & \text{with probability } 1 - U^\star_{(y+1)} \\ \text{uniform}(0,1) & \text{with probability } U^\star_{(y+1)} - U^\star_{(y)} \end{cases},$$

resulting in $\mathcal{R}_\theta(y) \sim \text{beta}(y+1, n-y+1)$. The fifth choice is simply to take the midpoint of the interval (i.e., $D = 0.5$). It is observed in simulation studies presented in Hannig (2009) that the second choice is optimal in some sense, though, there is a lack of intuition for why it seems to work better than simply taking the midpoint of the interval between the auxiliary endpoints.

6. Concluding remarks

Systematic investigation of pignistic transformations is critical to the broader understanding and appeal of imprecise probabilistic reasoning to mainstream statistical and machine learning communities, which are mostly familiar only with precise probabilistic reasoning derived from the Kolmogorov axioms. As stated previously, however, pignistic transformations have not received much attention, even in the imprecise probability literature. One notable exception is the brief Section 3.2 of Dubois et al. (2004) where such transformations are considered as a converse to the main results of the paper, and the proposed transformation relies heavily on a unimodality assumption. Another example is Smets (1990), relying on the insufficient reason principle. There remain many open research questions concerning pignistic transformations, particularly from a statistical inference perspective. Exploring optimization strategies—as motivated in this article by Theorem 18—in a new research area of *calculus of variations on credal sets* could be fruitful for new constructions of credal sets and pignistic transformations guided by finite-sample frequentist validity properties.

Acknowledgments

This manuscript benefited substantially from numerous discussions and discourse with Ryan Martin, Peter Grünwald, and one anonymous reviewer.

Appendix A. Probability dilation

One final remark on the topic of reliability is in order. An important issue is raised about the phenomenon of *dilation* in Grünwald (2018) concerning conditional validity as it relates to imprecise probability; see also Seidenfeld and Wasserman (1993). Consider the following simple illustration modified from Grünwald (2018). Suppose it is of interest to predict the event $\{U = 1\}$ for some random variable $U \sim \text{Bernoulli}(0.4)$. In this case $P(U = 1) = 0.4$. Now assume further that U depends in some unknown way on some other random variable $V \sim \text{Bernoulli}(0.5)$ that is observed. Without knowing how U depends on V, an imprecise probabilistic assessment might model the uncertainty by including all probability distributions P on $\{0,1\} \times \{0,1\}$ that satisfy:

$$0.4 = P(U = 1) = P(U = 1 \mid V = 0) \cdot 0.5 + P(U = 1 \mid V = 1) \cdot 0.5.$$

Consequently, the imprecise probabilistic assessment yields $P(U = 1 \mid V = v) \in [0, 0.8]$ regardless of the observed value $v \in \{0,1\}$. This means that after having gained information about V, we actually know less about the event $\{U = 1\}$ than we did ignoring V, i.e., information about $\{U = 1\}$ is lost. This phenomenon of dilation is indeed perplexing

because the obvious remedy of ignoring V conflicts with the precept that information should never be useless. I offer, however, the reconciliation that while information should never be useless, we must know how to use it for it to be useful; otherwise, it should not be used, as in the illustration.

Of course, imprecise probabilities are not necessarily a consequence of information lost due to dilation. The developments in this article illustrate how imprecise probabilities are a natural consequence of frequentist calibration.

Appendix B. Proof of Theorem 4

Proof Any realization of nonconformity scores $t_1(y_1), \ldots, t_{n+1}(y_{n+1})$ can be described as a collection containing unique values, $a_1, \ldots, a_K$, for some $K \leq n+1$ and occurring with frequencies $n_1, \ldots, n_K$, respectively (with $\sum_{k=1}^{K} n_k = n+1$). Using the fact that $t_1(Y_1), \ldots, t_{n+1}(Y_{n+1})$ are exchangeable (because $Y_1, \ldots, Y_{n+1}$ are exchangeable) as in Definition 2, it follows by definition that any realization $t_1(y_1), \ldots, t_{n+1}(y_{n+1})$ can be understood as some permutation of the values in the *bag* (i.e., a collection of elements with no ordering),

$$B := \{\!| \underbrace{a_{(1)}, \ldots, a_{(1)}}_{n_1}, \underbrace{a_{(2)}, \ldots, a_{(2)}}_{n_2}, \ldots, \underbrace{a_{(K)}, \ldots, a_{(K)}}_{n_K} |\!\},$$

where $a_{(k)}$ is the k-th order statistic (in ascending order) of the values $a_1, \ldots, a_K$. As such, the observed nonconformity scores $t_1(y_1), \ldots, t_{n+1}(y_{n+1})$ are just one of $(n+1)!$ equally possible permutations that could have been recorded, assuming y_{n+1} was generated from equation (1).

Next, with reference to the bag B, it can be determined that

$$\sum_{i=1}^{n+1} 1\{t_i(y_i) \geq t_{n+1}(y_{n+1})\} = \begin{cases} n+1 & \text{if } \; t_{n+1}(y_{n+1}) = a_{(1)} \\ n+1-n_1 & \text{if } \; t_{n+1}(y_{n+1}) = a_{(2)} \\ n+1-n_1-n_2 & \text{if } \; t_{n+1}(y_{n+1}) = a_{(3)} \\ & \vdots \\ n_K & \text{if } \; t_{n+1}(y_{n+1}) = a_{(K)} \end{cases}.$$

Furthermore, there are $n! \cdot n_j$ permutations of the values in B in which the last reported value, $t_{n+1}(y_{n+1}) = a_{(j)}$, so it must be the case that

$$P\left(\sum_{i=1}^{n+1} 1\{t_i(Y_i) \geq t_{n+1}(Y_{n+1})\} = v \;\middle|\; B \right) = \begin{cases} \frac{n! \cdot n_1}{(n+1)!} = \frac{n_1}{n+1} & \text{if } \; v = n+1 \\ \frac{n! \cdot n_2}{(n+1)!} = \frac{n_2}{n+1} & \text{if } \; v = n+1-n_1 \\ \frac{n! \cdot n_3}{(n+1)!} = \frac{n_3}{n+1} & \text{if } \; v = n+1-n_1-n_2 \\ & \vdots \\ \frac{n! \cdot n_K}{(n+1)!} = \frac{n_K}{n+1} & \text{if } \; v = n_K \\ 0 & \text{else} \end{cases}.$$

Note that in the special case without repeated values (i.e., $n_1 = \cdots = n_K = 1$), the above expression reduces to a discrete uniform probability mass function. In any case,

$$\begin{aligned} P(\Gamma_n^\alpha \not\ni Y_{n+1} \mid B) &= P\{p_n(Y_{n+1}) \leq \alpha \mid B\} \\ &= P\bigg(\sum_{i=1}^{n+1} 1\{t_i(Y_i) \geq t_{n+1}(Y_{n+1})\} \leq \alpha(n+1) \;\bigg|\; B \bigg) \\ &= \begin{cases} \frac{n_K}{n+1} + \frac{n_{K-1}}{n+1} + \cdots + \frac{n_{k_\alpha+1}}{n+1} & \text{if } k_\alpha < K \\ 0 & \text{if } k_\alpha = K \end{cases}, \end{aligned}$$

where $k_\alpha := \min\big\{j \in \{0, \ldots, K\} \,:\, n + 1 - \sum_{k=0}^{j} n_k \leq \alpha(n+1)\big\}$ and $n_0 := 0$. Observe from the construction of k_α that,

$$n_K + n_{K-1} + \cdots + n_{k_\alpha+1} = n + 1 - \sum_{k=0}^{k_\alpha} n_k \leq \alpha(n+1),$$

and divide by $n+1$ on all sides. Thus, in any case,

$$P(\Gamma_n^\alpha \not\ni Y_{n+1}) = \int P(\Gamma_n^\alpha \not\ni Y_{n+1} \mid B) d\nu(B) \leq \alpha \cdot \int d\nu(B) = \alpha,$$

where $\nu(\cdot)$ is any probability measure describing the uncertainty in observing the bag B. ■

References

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2023.

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19 (2):1641–1662, 2025.

T. Augustin and F. P. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2):251–272, 2004.

M. S. Balch, R. Martin, and S. Ferson. Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A*, 475(20180565), 2019.

O. Basir and X. Yuan. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. *Information Fusion*, 8(4):379–386, 2007.

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

I. Bloch. Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recognition Letters*, 17(8):905–919, 1996.

I. Carmichael and J. P. Williams. An exposition of the false confidence theorem. *Stat*, 7(1): e201, 2018.

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

F. P. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36(4):349–357, 1998.

F. P. Coolen and T. Coolen-Maturi. Nonparametric predictive inference. In *International Encyclopedia of Statistical Science (2nd Edition)*. Springer, 2024.

A. P. Dempster. On direct probabilities. *Journal of the Royal Statistical Society: Series B*, 25(1):100–110, 1963.

A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37(2):355–374, 1966.

T. Denoeux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 737–760. Springer, 2008.

N. Dey and J. P. Williams. Valid inference for machine learning model parameters. *arXiv preprint arXiv:2302.10840*, 2023.

N. Dey, R. Martin, and J. P. Williams. Generalized universal inference on risk minimizers. *Journal of the Royal Statistical Society: Series B*, pages 1–23, 2025. doi: 10.1093/jrsssb/qkaf065.

L. Díaz-Más, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer. Shape from silhouette using Dempster-Shafer theory. *Pattern Recognition*, 43(6):2119–2131, 2010.

D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10(4):273–297, 2004.

R. A. Fisher. Inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 528–535. Cambridge University Press, 1930.

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

I. Gelfand and S. Fomin. *Calculus of variations, (Translated and edited by Silverman, RA)*. Dover Publications, 2000.

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

P. Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024.

J. Hannig. On generalized fiducial inference. *Statistica Sinica*, 19(2):491–544, 2009.

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

B. M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691, 1968.

B. M. Hill. De Finetti's theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion). *Bayesian Statistics*, 3:211–241, 1988.

A. Hjort, G. H. Hermansen, J. Pensar, and J. P. Williams. Uncertainty quantification in automated valuation models with spatially weighted conformal prediction. *International Journal of Data Science and Analytics*, pages 1–18, 2025.

P. E. Jacob, R. Gong, P. T. Edlefsen, and A. P. Dempster. A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116(535): 1181–1192, 2021.

H. Jeffreys. On the theory of errors and least squares. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 138 (834):48–55, 1932.

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

R. Martin and C. Liu. *Inferential models: Reasoning with uncertainty*, volume 145. CRC Press, 2015.

B. Park, S. Balakrishnan, and L. Wasserman. Robust universal inference for misspecified models. *Biometrika*, 113(2):asaf070, 2026.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

T. Schweder and N. L. Hjort. *Confidence, likelihood, probability*, volume 41. Cambridge University Press, 2016.

T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *Annals of Statistics*, 21 (3):1139–1154, 1993.

G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

G. Shafer. Comment on "A Gibbs sampler for a class of random convex polytopes," by Pierre E. Jacob, Ruobin Gong, Paul T. Edlefsen, and Arthur P. Dempster. *Journal of the American Statistical Association*, 116(535):1196–1197, 2021.

P. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 10 of *Machine Intelligence and Pattern Recognition*, pages 29–39. North-Holland, 1990.

J. Tie, J. Cohen, K. Lahouel, S. Lo, Y. Wang, S. Kosmider, R. Wong, J. Shapiro, M. Lee, S. Harris, A. Khattak, M. Burge, M. Harris, J. Lynam, L. Nott, F. Day, T. Hayes, S. McLachlan, B. Lee, J. Ptak, N. Silliman, L. Dobbyn, M. Popoli, R. Hruban, A. Lennon, N. Papadopoulos, K. Kinzler, B. Vogelstein, C. Tomasetti, and P. Gibbs. Circulating tumor DNA analysis guiding adjuvant therapy in stage II colon cancer. *New England Journal of Medicine*, 386(24):2261–2272, 2022.

P. Vasseur, C. Pégard, E. Mouaddib, and L. Delahoche. Perceptual organization approach based on Dempster-Shafer theory. *Pattern Recognition*, 32(8):1449–1462, 1999.

V. Vovk. Universal predictive systems. *Pattern Recognition*, 126:108536, 2022.

V. Vovk. Conformal predictive distributions: An approach to nonparametric fiducial prediction. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 364–380. Chapman and Hall/CRC, 2024.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

V. Vovk, J. Shen, V. Manokhin, and M. Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474, 2019.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

J. P. Williams. Discussion of "A Gibbs sampler for a class of random convex polytopes". *Journal of the American Statistical Association*, 116(535):1198–1200, 2021.

J. P. Williams and Y. Liu. Decision theory via model-free generalized fiducial inference. In *Belief Functions: Theory and Applications*, volume 14909, pages 131–139. Springer, 2024.

M. Xie and K. Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.