So What are *Good* Imprecise Forecasts?[*]

Rajeev Verma[†]
UvA-Bosch Delta Lab
University of Amsterdam
r.verma@uva.nl

Rabanus Derr[†]
Foundations of Machine Learning Systems
University of Tuebingen
rabanus.derr@uni-tuebingen.de

Christian A. Naesseth
UvA-Bosch Delta Lab
University of Amsterdam

Volker Fischer
Bosch Center for Artificial Intelligence

Eric Nalisnick
Department of Computer Science
Johns Hopkins University

†

Abstract

We explore the place of imprecise forecasts in machine learning. While learning systems are increasingly used to inform downstream decisions, the standard framework of expected utility and calibrated probability forecasts may be insufficient when data arises from heterogeneous or ambiguous sources. We argue that the resulting institutional separation between learning and decision-making motivates a broader notion of forecasting, one that can accommodate ambiguity without assuming a single underlying distribution. Through analysis of the multi-distribution learning framework, we show that loss-minimizing predictors can fail to provide useful decision recommendations or actuarial decisions, calling for the adoption of imprecise forecasts. Finally, we propose a statistical mechanism to evaluate such forecasts by *testing by betting*, offering a falsifiable analogue of calibration. Our goal is to ground imprecision as a necessary and testable feature of epistemically sound machine learning.

1 Introduction

The field of machine learning is largely under the grip of expected utility theory (EUT) [Bernoulli, 1738, von Neumann and Morgenstern, 1944, Savage, 1954] (and probability theory as a consequence). Originally introduced by Bernoulli, the *utility*—an invisible quantity and subjective to decision-maker—is combined with *probabilities* over uncertain outcomes to choose among rational decisions. However, EUT is not the most accepted theory of decision-making, and there have always been arguments against it, particularly highlighting its inability to account for systematic behavioral phenomena like ambiguity aversion [Camerer and Weber, 1992]. Some scholars argue that *imprecise probabilities* can provide a solution. As the name suggests, imprecise probability means a non-empty set of probability measures $\mathcal{Q}$ with an associated non-linear expectation functional $\bar{\mathbb{E}}_{\mathcal{Q}}\left[\cdot\right] := \sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q}\left[\cdot\right]$, and it has been motivated from several perspectives [Artzner et al., 1999]. Most notably, Gilboa and Schmeidler [1989] derived this decision-rule in an axiomatic setup (akin to EUT) under the setting of ambiguity as highlighted by Ellsberg [1962] as the most celebrated criticism against EUT.

Recently, machine learning scholars have begin to take a keen interest in imprecise probabilities [Caprio et al., 2024, Singh et al., 2024, Fröhlich and Williamson, 2024, Manchingal et al.,

[*] We thank Erik Quaeghebeur, Krikamol Muandet, and Anurag Singh for helpful discussions that helped clarify the setup of this work.

[†] Equal contribution.

1st Workshop on Epistemic Intelligence in Machine Learning (EIML 2025) .

2025, Singh et al., 2025]. Imprecise probabilities promise *epistemic humility*: a framework to express and reason under what one does not know, and allows one to accommodate ambiguity and indecision—some of the important capabilities in a truly intelligent agent [Russell, 2019]. However, there are many connotations to the word *epistemic*, and while it would be good to have *epistemic humility*, it is not yet clear it is ever warranted. The greatest challenge to exploiting the promise of imprecise probability, we argue, is the meaning and the evaluation of it. This challenge further exacerbates in a data-driven manner machine learning works. Contrary to classical statistics where the focus is on to estimate "true" parameters of the phenomena of interest, machine learning side steps this by giving estimates that "work well". For good decisions, one does not necessarily need "true" probabilities, instead the calibrated ones are good enough to drive *actuarial decisions*. Furthermore, there is not a good evaluation mechanism to check if one has estimated true probabilities, one can only run some statistical tests to falsify the wrong estimates—calibration of forecasts being one of such tests. Most recently, Singh et al. [2025] circumvent some challenges in this direction by proposing the first proper scoring rule to incentivize truthful elicitation of imprecise forecasts, i.e., truthfully reporting an imprecise forecasts minimizes the expected proper score. However, they take subjectivist perspective for the meaning of imprecision, and it is not immediately clear what direct objective translation can be attached to our setting of data-driven machine learning. Furthermore, our position is that elicitation is not evaluation, and the challenge to build a statistically viable mechanism to evaluate imprecise forecasts (akin to the calibration) stands still.

Summary of our contributions. In this *preliminary* study, we address two questions: what is the need for imprecise forecasts in standard data-driven machine learning setting? Standard prediction systems do not aim to estimate the "true" probabilities, good calibrated probabilities can be used to drive decisions via the expected utility theory. In this section, we offer theoretical results why that might be insufficient for practical learning scenarios, calling to adopt the imprecision in forecasts. As a second question, we ask how can one statistically falsify imprecise forecasts. This would enable one to evaluate imprecise forecasts the way precise forecasts are evaluated via their calibration. To this end, we propose to adopt the *testing by betting* framework [Ramdas et al., 2023].

2 Preliminaries

In this section, we lay down the formulation we follow in this work. We adopt upright lettering $(\mathbf{x}, \mathrm{y}, \mathbf{z})$ to denote random variables, where boldface letters denote multi-dimensional objects.

Learning, decisions, and institutional separation. We consider the standard setup of classification. In particular, we have the space $\Omega := \mathcal{X} \times \mathcal{Y}$, with the assumption of $|\Omega| = \kappa \in \mathbb{N}$, and for simplicity, we assume $\mathcal{Y} = \{0, 1\}$. Here, $\boldsymbol{x} \in \mathcal{X}$ denotes the input, and $y \in \mathcal{Y}$ denotes the associated label. A practical machine learning problem starts with some given observations from this space as a dataset $D := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. The goal of *learning* is to find a (confidence) predictor function $h\colon \mathcal{X} \to \Delta$ for a certain loss function $\ell\colon \mathcal{Y} \times \Delta \to [0, 1]$. Given $\mathcal{H}$ as the set of functions of the above type, the said predictor is obtained as $h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell\left(\mathrm{y}_i, h\left(\boldsymbol{x}_i\right)\right)$[2]. Furthermore, *assuming* that there is some distribution P on Ω, and D is sampled in an *i.i.d.* fashion from P, we have $h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, \mathrm{y}) \sim P}\left[\ell\left(\mathrm{y}, h\left(\mathbf{x}\right)\right)\right]$.

Given h as the learned predictor, a *decision-problem* concerns with translating predictions to decisions. Given a downstream decision-maker with finite action space $\mathcal{A}$ and a bounded (dis)utility function $c\colon \mathcal{Y} \times \mathcal{A} \to [0, 1]$, and the goal is to use the h to recommend decisions as $\mathrm{BR}(\boldsymbol{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\mathrm{y} \sim h(\boldsymbol{x})}\left[c\left(\mathrm{y}, a\right)\right]$[3].

As an illustrative example of our distinction of *learning* and *decisions*, consider an example of a predictive diagnosis system to help diagnose whether some CT-scan imagery denotes pneumonia or not. A learning problem will concern itself with training such a predictive system with the goal of good accuracy. A decision-problem, however, is a deployment scenario where a medical professional will use this predictive system to allocate arbitrary decisions—to treat the patient or not, to recommend a certain treatment or another, etc. This distinction further motivates our next definition:

[2] Here, we assume such predictor h^* can be uniquely identified to simplify the analysis.

[3] We denote the tie-breaking rule with δ. It maps a set $A \in \mathcal{A}$ to an element $a \in A$.

Definition 2.1. *(Institutional separation).* Institutional separation refers to lack of information for the learner about the deployment time decision-problem.

We argue that this is a practical setting in modern machine learning and AI systems. An enterprise trains (or learns) a general purpose predictive system to be used by arbitrary decision-makers, and there could be multiple decision-makers using the same learned predictor. Additionally, utilities of the same decision-maker could arbitrarily change. For example, the same medical professional could be risk averse on situations of limited resources, and could be risk-taking in other scenarios. As a result of this mis-alignment of objectives, a favorable hypothesis (or action) as per the learning problem cannot be necessarily used to derive favorable actions in the decision-problem. Thus, in the case of institutional separation, the learning problem comes with an additional "burden" of learning in a way so as to bridge this institutional separation.

Institutional separation, calibration, and actuarial decisions. Institutional separation requires one to put additional evaluation criteria on the learning problem, beyond the standard "minimal" loss criterion. This evaluation criterion has been previously motivated from the perspective of *actuarial decisions*, i.e. predictions should allow the decision-makers to correctly gauge the consequences of their actions [Seidenfeld, 1985, Schervish, 1989, Grünwald, 2018, Sahoo et al., 2021, Zhao et al., 2021, Fröhlich and Williamson, 2024, Derr et al., 2025]. The standard result in this literature is that the calibrated predictor h bridges the institutional separation, as stated below:

Definition 2.2. *(Calibration and actuarial decisions).* For $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{0, 1\}$ and a distribution P on it, with marginal $P_{\mathbf{x}}$ and denoting $\eta\left(\boldsymbol{x}\right) := P\left(\mathrm{y} = 1 \mid \mathbf{x} = \boldsymbol{x}\right)$, a predictor $h\colon \mathcal{X} \to \Delta$ is called *calibrated* if

$$\mathbb{E}\left[\eta\left(\mathbf{x}\right) \mid h\left(\mathbf{x}\right) = p\right] = p, \ \ \forall p.$$

Next, for the decision-problem with action set $\mathcal{A}$, (dis)utility function $c\colon \mathcal{Y} \times \mathcal{A} \to [0, 1]$, a decision-maker with the best-response policy given the predictor h, i.e., $BR\left(\boldsymbol{x}\right) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\mathrm{y} \sim h(\boldsymbol{x})}\left[c\left(\mathrm{y}, a\right)\right]$ and tie-breaking rule δ, then the following holds:

$$\mathbb{E}_{\mathbf{x}, \mathrm{y} \sim P}\left[c\left(\mathrm{y}, \delta(BR\left(\mathbf{x}\right))\right) - \mathbb{E}_{\mathrm{y} \sim h(\mathbf{x})}\left[c\left(\mathrm{y}, \delta(BR\left(\mathbf{x}\right))\right)\right]\right] = 0.$$

Informally, the above expression says that the best-responding decision-maker to the calibrated forecast would not incur *sure-loss*—or calibrated predictor provides actuarial decisions. Calibration of the predictor roughly means the alignment of predictions and actual occurrence of outcome on average, and these two properties can be said to imply that calibration induces an indistinguishability criterion [Gopalan and Hu, 2025]. Furthermore, calibration can also be used as a statistical test for falsification: a predictor that is not calibrated is not a *good* predictor, but the converse is not true. For example, the marginal predictor $\mathbb{E}[\mathrm{y}]$ is calibrated, however it is relatively uninformative. Multi-calibration [Hebert-Johnson et al., 2018] strengthens the promise of regular calibration, and the related framework of omniprediction [Gopalan et al., 2021] adds on to that by promising stronger notion of actuarial decisions. It seems that if one is to care about actuarial decisions, then calibrated predictors is all one need. However, we bring attention to the assumption of the data D being sampled in an *i.i.d.* manner, and argue that in several scenarios, that is not a realistic assumptions.

Multi-distribution learning (MDL). Modern challenges of robustness [Kuhn et al., 2025], privacy [Mohri et al., 2019], fairness [Martinez et al., 2020], and reliability of machine related decisions has led to the development of the MDL framework [Haghtalab et al., 2022] as an extension of the celebrated agnostic learning setup [Kearns et al., 1992], where the data comes from a *set of distributions*. Of particular interest to us is fairness: powered by the promise of data and learning, policy makers and private institutions are increasingly relying on automated decision-support systems to allocate critical resources [Perdomo, 2024] in user-facing scenarios like welfare [Vaithianathan et al., 2021, Ahn et al., 2024]. However, as the pervasiveness of predictions increase in defining and structuring social outcomes, it is crucial to be cognizant of the *disparate impact* they can cause, as has been empirically observed. Fairness requires the predictions to not be systematically biased across heterogeneous subpopulations defined by some protected attributes, and MDL can also instantiate this issue by considering these heterogeneous populations as the set of interest.

Formally, we denote $\mathcal{Q} := \{Q_1, Q_2, \ldots, Q_\eta\}$ as a set of distributions on $\Omega := \mathcal{X} \times \mathcal{Y}$. Given a loss function $\ell\colon \mathcal{Y} \times \Delta \to \mathbb{R}$, multi-distribution learning tries to solve the optimization problem, $\arg\min_{h\colon \mathcal{X} \to \Delta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$. We will discuss the shortcomings of this approach in Section 3.

Decision behavior facing multiple distributions. Let us take a short detour to set light on the background of the minmax-objective in MDL as a certain choice of decision behavior. For the sake of simplicity, consider a set of distributions $\mathcal{Q}$ only on $\mathcal{Y}$. Let there be a decision-maker with finite action set $\mathcal{A}$ and (dis)utility function $c\colon \mathcal{Y} \times \mathcal{A} \to [0,1]$. We define $R_c\left(Q,a\right) := \mathbb{E}_Q\left[c\left(\mathrm{y},a\right)\right]$ for a single $Q \in \mathcal{Q}$. In this case, a decision-maker will choose an action as $a^* = \arg\min_{a\in\mathcal{A}} R_c\left(Q,a\right).$ What decision-making criteria will the decision-maker apply in the case of multiple distributions? A standard criterion in decision-theory is that of admissibility, defined below:

Definition 2.3. *(Admissibility and Dominance).* An action $a \in \mathcal{A}$ is said to be admissible with respect to (w.r.t.) $\mathcal{Q}$ for the loss function c, if there exists no such $a' \in \mathcal{A}$ such that the following holds:

$$R_c(Q,a') \leq R_c(Q,a)\ \forall Q \in \mathcal{Q} \text{ and } \exists Q_0 \in \mathcal{Q} \text{ s.t. } R_c(Q_0,a') < R_c(Q_0,a),$$

and in case such an a' exists, we call a' to dominate a.

Thus, admissible actions are not dominated. Informally, for decision-making, not considering a non-admissible action means ruling out the action for which there certainly exists an action that is at-least as better this action, and hence intuitively does not make sense for the decision-maker to ever consider a non-admissible action. However, the set of admissible actions could be large, and while it is useful in ruling out "bad" actions, it does not further tell the decision-maker how to choose among the admissible actions. To this end, one can rely on the following classical result in decision-theory:

Proposition 2.4. *from* [Hoff, 2013]. *Any admissible action a w.r.t. the finite $\mathcal{Q}$ can be recovered as a result of the problem* $\arg\min_{a\in\mathcal{A}} \int R_c\left(Q,a\right)\pi\left(dQ\right)$*, the latter is also referred to as the Bayes action for some prior π on $\mathcal{Q}$.*[4]

Thus, if the admissibility criterion rules out the trivially "bad" actions, the very act of choosing one action among the admissible ones resorts to putting a prior (or a second-order distribution) over the uncertainty set $\mathcal{Q}$. Informally, this prior could encapsulate the beliefs of the decision-maker across $\mathcal{Q}$ to weigh the sources of uncertainty. A very standard approach is to choose some prior π. However, it trivially follows from the definition of admissibility that such a decision procedure would cause disparate impact across different distributions. In particular, if we understand standard empirical risk minimization (ERM) as a decision problem, i.e., the action set $\mathcal{A}$ is the set of hypothesis $\{h\colon \mathcal{X} \to \Delta\}$, the cost function c is a loss function ℓ, and the base set $\mathcal{Y}$ is extended to $\mathcal{X}\times\mathcal{Y}$, then we can look on the widely documented disparate impact of machine learning models from a different angle. Our formulation reveals its fundamental nature—ERM across heterogeneous sub-populations is fundamentally *unfair*, independent of the finite sample and approximate optimization issues. Another approach to choose an admissible action is based on the axiomatic framework by [Gilboa and Schmeidler, 1989].

$$a^* = \min_{a\in\mathcal{A}} \sup_{\pi} \int R_c\left(Q,a\right)\pi\left(dQ\right).$$

The corresponding learning problem is the standard formulation of the MDL problem studied by Haghtalab et al. [2022] and other subsequent works in the learning theory literature. It directly controls the worst-case loss, and forms a natural choice in the case of institutional separation we have. In the next section, we will show that this framework does not provide actuarial decisions, leading to the necessity of imprecise forecasts.

3 Multi-Distribution Learning and Why Precise Forecasts Are Not Enough

In the previous section, we motivate MDL as the natural framework to address challenges of robustness, reliability, and fairness, in particular. In this section, we investigate MDL under the standard setup of *learning* and *decisions*, and question if the framework can actually provide generally useful recommendations for decisions or actuarial decisions. Our results affirmatively say no. In a way, we echo the findings of Verma et al. [2025] who argued that a *perfect* predictor obtained as a result of loss minimization cannot be guaranteed to be calibrated across distributions in the set $\mathcal{Q}$.

Note that under some regularity conditions, we can ensure the existence of a *perfect predictor* in multi-distribution learning setup, $h_\ell = \arg\min_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y},h(\mathbf{x}))]$ (see Lemma A.2). In particular, if ℓ is strictly proper, then $h_\ell = \bar{Q}_{\mathrm{y}|\mathbf{x}} = \bar{Q}(\cdot|\mathbf{x})$ for some $\bar{Q} \in$

[4] We note the finiteness of $\mathcal{Q}$, $\mathcal{A}$, Ω, and the boundedness of c to be able to apply this result.

$\arg\max_{Q\in\mathcal{Q}} \inf_{h\colon \mathbf{x}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$, i.e., the perfect predictor is equal to the conditional distribution of the distribution $\bar{Q}$. Similar result have appeared in Fröhlich and Williamson [2024], Verma et al. [2025], Rahimian and Mehrotra [2022], Kuhn et al. [2025], and the argument builds on the heavy usage of the results in Grünwald and Dawid [2004].

Loss minimization. Our next statement establishes the pathology of a *perfect* predictor in a multi-distribution setup for arbitrary decision-maker.

Proposition 3.1. (MDL predictors are (strictly) dominated). *For a learning problem with a loss function $\ell\colon \mathcal{Y}\times\Delta \to [0,1]$, the set $\mathcal{Q}$ of distributions with the perfect predictor $h_\ell \in \arg\min_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$, for a decision problem (dis)utility function $c\colon \mathcal{Y}\times\mathcal{A}\to[0,1]$, finite action set $\mathcal{A}$, and best response mechanism BR, Then, there exists $h_c\colon \mathcal{X}\to\Delta$, such that, $\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x}))] \geq \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))]$. Furthermore, for every strictly proper loss function ℓ, there exists a closed convex set of distributions $\mathcal{Q}$ and a decision maker with (dis)utility function c, such that the inequality is strict. (Proof in Appendix A.1 and Proposition A.1)*

The above argument states that any perfect predictor recommends decisions to decision makers which are weakly dominated, in terms of worst-case expected utility, by the decisions recommended by a predictor which is tailored to the needs of the decision maker. More problematically, we show that for every general purpose loss function ℓ there exists an ambiguity set $\mathcal{Q}$ and a decision maker, such that the recommended decisions are *strictly* dominated, in terms of worst-case expected utility, by the decisions recommended by a predictor which is tailored to the needs of the decision maker. In conclusion, perfect predictors on ambiguity sets can be (strictly) suboptimal in recommending actions to decision makers when the general purpose loss function for learning and the dis-utility of the decision makers are mis-aligned.

Calibration and actuarial decisions. Earlier, we argued that to truly bridge the institutional separation, one needs to provide actuarial decisions, i.e. the decision-makers should be able to accurately estimate the consequences of their actions via accurate (dis)utility estimation. It is known that a perfect MDL predictor does not satisfy this requirement (Proposition A.3 in Appendix). Specifically, the perfect predictor for a general purpose loss function on an ambiguity set $\mathcal{Q}$ does not need to provide actuarial insurance to an aligned decision maker. That is, the expected loss with respect to the general purpose loss function estimated by using the probabilistic forecast, is *not* necessarily lower bounded by the actual incurred loss for some distribution in $\mathcal{Q}$. Actuarial decisions are promised by calibration, i.e., given a predictor h_ℓ, if it holds that $\mathbb{E}_Q\left[\mathrm{y}|\, h_\ell\left(\mathbf{x}\right) = p\right] = p,\ \forall p \in \{h_\ell\left(\mathbf{x}\right), \boldsymbol{x}\in\mathcal{X}\},\ Q\in\mathcal{Q}$. Verma et al. [2025] argued that a perfect MDL predictor as a result of loss minimization is not guaranteed to be calibrated across $\mathcal{Q}$. However, their result does not rule out that one cannot post-process it to calibrate it for each $Q\in\mathcal{Q}$ (in the same fashion as multi-calibration). We show that post-hoc calibration is strongly limited.

Proposition 3.2. *Given $\mathcal{X}\times\{0,1\}$ be the finite input-output space. Consider two distributions Q and Q' on it such that $Q_{\mathbf{x}} = Q'_{\mathbf{x}}$ and $\mathbb{E}_Q[\mathrm{y}] \neq \mathbb{E}_{Q'}[\mathrm{y}]$, then a single precise predictor $h : \mathcal{X}\to[0,1]$ cannot be calibrated for both Q and Q'. (Proof in Appendix B.1).*

Consider the case for fairness argued above. Given a predictor as a result of a learning problem, heterogeneous populations would want to be able to use this predictor to drive actuarial decisions. However, our results indicate that an MDL predictor, when the sets of distributions involve heterogeneous populations, can lead to disparate impact. We argue in Appendix B that the situation is even worse when legal requirements forbid the usage of protected attributes.

4 A Final Illustrative Example for the Necessity of Imprecise Forecasts

In the previous section, we highlighted two facets of shortcomings of precise (perfect) forecasts in multi-distribution learning. Let us further illustrate the necessity of imprecise forecasts via a simple example. Consider two coins with biases 0.4 and 0.6 which are tossed alternately, resulting in a long sequence of outcomes of coin flips. Imagine some forecaster is asked to give the estimate of the probability of heads as per the resulting outcome sequence. Since there will be equal proportions of heads and tails in the outcome sequence, a reasonable forecast is 0.5. Does this forecast provide actuarial decisions? Without referring to any specific decision problem, we can consider the gambling / betting interpretation of probability. Under this betting view, the quoted "probability" is simply the price of a \$1 ticket that pays if the next outcome is heads. If the forecaster always posts the price 0.5, an active bettor who knows which coin will be tossed can take the other side whenever the price is off. On turns using the 0.6 coin, the bettor buys the heads ticket; on turns using the 0.4 coin, the bettor sells it (or, equivalently, buys a tails ticket). In each case the bettor is trading against a mis-priced ticket and earns a small, repeatable edge per dollar staked. Over many flips this produces a systematic gain, revealing that the 0.5 forecast is not suitable for guiding decisions even though it matches the long-run frequency of heads. Or the forecaster has exposed themselves to the sure-loss.

The outcome sequence generated as a result of the coin flips has predictable regularity that the bettor is able to exploit. Formally, there exists a subsequence of outcomes such that the outcome rate does not match the forecast, i.e. the outcome sequence follows statistical stability [Gorban, 2017] w.r.t. two biases (0.4 and 0.6), and the forecaster is not aware of it. Equivalently, the outcome sequence is not generated in an *i.i.d.* fashion. There are two ways out: first, we reason to give the imprecise forecast $[0.4, 0.6]$, or second, we argue to condition the forecast on the coin information. The second requires a certain level of clairvoyance as it asks: the forecaster to forecast 0.6 when the coin to be tossed is one with the bias 0.6, and 0.4 otherwise. Hence, for a realistic forecaster who cannot afford this clairvoyance, imprecise forecasts become necessary to avoid sure-loss. We reflect further on this in Appendix B.

5 Evaluation of Imprecise Forecasts

So far, we have argued for the pitfalls of regular loss minimization when the data comes from a set of distributions. Then we showed that the standard formalism of MDL is not guaranteed to provide actuarial decisions. The issue becomes glaring in situations where data inherently comes from a heterogeneous set of populations (which it does), and one is not allowed access to heterogeneity markers, there is a natural risk for automated decision-making to cause fundamental disparate impact. Our position is that imprecise forecasts can help overcome it, and we reason a practical demonstration is need to further argue for it.

Singh et al. [2025] proposed the scoring rule to elicit imprecise forecasts. However, to be able to fully exploit the promise of imprecise forecasts, we need to build a statistical way to test the forecasts. Consider an agent who gives us the forecast for some event to be in the interval $[0.4, 0.6]$, how one can test its correctness? We look up to Popperian falsifiability [Popper, 1959], and argue for the mechanism to execute it. Precise forecasts are falsifiable by calibration—if the forecaster's forecast do not match the average outcome rate, one can falsify the forecaster. Standard calibration metrics like the expected calibration error (ECE) [Naeini et al., 2015], smooth calibration error (smECE) [Foster and Hart, 2018, Błasiok et al., 2023] then quantify the mis-alignment of the forecast and the outcome. Alternatively, one can also operationalize this alignment via the *no sure-loss* property, i.e. checking how actuarially aligned the forecasts are to the eventual decisions. Such a formulation was also used by Fröhlich and Williamson [2024] to define the calibration of imprecise forecasts, albeit in relation to some utility function and a specific min-max decision-rule. Independent of any utility function or the decision-rule, we can generalize this formulation adopting the betting rate interpretation of the probability, and embed the calibration of imprecise forecasts in the recently popularized framework of *testing by betting* [Ramdas et al., 2023].

Testing by betting of imprecise forecasts. For simplicity, we consider the setting $\mathcal{X} \times \{0, 1\}$, and hence the imprecise forecast is for the outcome $\mathrm{y} = 1$. Give the imprecise forecast $[a, b]$, we offer the following two test-statistics for evaluating imprecise forecasts:

$$M_t^U := \prod_{i=1}^{t} \left(1 + A_i \lambda_i \left(\mathrm{y}_i - b\right)\right), \quad \text{and} \quad M_t^L := \prod_{i=1}^{t} (1 + A_i \zeta_i \left(a - \mathrm{y}_i\right),$$

where $A_i \in \{0, 1\}$ is a predictable selector to decide whether to bet on round i or not, and allows for adaptive subsequence search, $\lambda_i \in [0, \frac{1}{b})$ and $\zeta_i \in [0, \frac{1}{1-a})$ represent the *betting rates* chosen adaptively by the bettor at each round to maximize their expected wealth under the announced imprecise forecast. Intuitively, the bettor starts with unit wealth and sequentially bets a fraction λ_i of their current wealth against the upper endpoint b (or ζ_i against the lower endpoint a). Under the null that the forecast intervals are statistically valid, the resulting wealth processes $(M_t^U)_{t\geq 1}$ and $(M_t^L)_{t\geq 1}$ form *non-negative supermartingales*, ensuring $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}$. Consequently, any systematic increase in wealth provides valid evidence against the forecaster's correctness. In other words, testing by betting converts the notion of miscalibration into a tangible betting gain: if a bettor can grow wealth unboundedly, the forecast must be statistically invalid. We provide further details in Appendix C.

This testing procedure also relates to the notion of imprecise randomness as studied by de Cooman and De Bock [2022]—we are testing for the alignment of the forecasts and the outcomes that *adhere* to the notion of imprecise randomness. We plan to elaborate the argument in future work. We assert that the stated evaluation mechanism is directly testing for the *no sure-loss* property of the forecasts in reference to the forecaster. However, imprecise forecasts come with an additional challenge: does the imprecise forecast contain the *right* amount of imprecision? For example, a vacuous predictor, a predictor predicting the full probability simplex, is correct as it avoids any sure-loss, however, it is also incorrect as it is unnecessarily imprecise. Thus, evaluating imprecise forecasts via either no sure-loss property or the size is incomplete—we need both criteria (Derr and Williamson [2025] refer to this problem as the *imprecision problem*).

The imprecision problem. While we do not have a complete answer to the imprecision problem, our position consists of two central claims. (a) The evaluation of imprecise forecasts can be understood as a useful certificate instead of an incentivizing objective. Calibrated forecasts can be simply achieved by correct base rate predictions. Furthermore, calibration is known to not incentivize truthful forecasting [Seidenfeld, 1985]. Nevertheless, calibration as evaluation criterion has proven itself useful [Derr et al., 2025]. Analogously, evaluation criteria for imprecise forecast might potentially not incentivize the "right" amount of imprecision, but still give valuable insights, respectively certify the quality of the forecasts. (b) Furthermore, the evaluation of the forecasts has to be contextualized with respect to who is using those forecasts. A forecast is good as long as it enables good decisions to the users. In the context of the testing procedure outlined above, a forecast is good if no user one can exploit the forecaster. Consider a set of decision makers defined as $\mathcal{G} = \{G_1, G_2, \dots, G_\alpha\}$, an imprecise forecast will have the right amount of imprecision if it prohibits collective exploitation by $\mathcal{G}$, and no one else is considered. Thus, our position is that the imprecision problem has to be relativized with respect to the users / evaluators, as "there are no (universally) good forecasts," as Verma [2024] notes.

6 Conclusions

Machine learning is about predictions, and these predictions are increasingly affecting the social structure of the society by informing decision-makers in resource allocation settings. Such predictions / forecasts are evaluated via (multi)calibration, and while there has been some concerns as calibrated predictors do not encode epistemic humility [Hullman, 2024], it is not clear what qualifies as good epistemic uncertainty. In this work, we contribute to this discussion by arguing the pitfalls of learning and decision-making in the case of multi-distribution learning where the data does not come in an *i.i.d.* manner from a single distribution, but from a set of distributions, and we show that imprecise forecasts can help provide actuarial decisions. On the goodness of epistemic uncertainty, our position is to quantify this heterogeneity in the dataset. Furthermore, we provide a statistical mechanism to test the imprecise forecasts via betting. However, this is a preliminary work, and our argument is far from finished. Our goal in the future is to further investigate what imprecision can contribute to the practice of machine learning by simulations—this involves quantifying and evaluating imprecise forecasts in a data-driven manner.

Acknowledgments and Disclosure of Funding

UvA-Bosch Delta Lab at the University of Amsterdam is funded by the Bosch Center for Artificial Intelligence. Thanks to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting RD. RD was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645; he was also supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center.

References

Eunhye Ahn, Ruopeng An, Melissa Jonson-Reid, and Lindsey Palmer. Leveraging machine learning for effective child maltreatment prevention: A case study of home visiting service assessments. *Child Abuse & Neglect*, 2024.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 1999.

Daniel Bernoulli. Specimen theoriae novae de mensura sortis. *Commeniarii Academiae Scientiarum Lmperialis Petropolitanae*, 1738.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks. *arXiv preprint arXiv:2304.09424*, 2023.

Colin Camerer and Martin Weber. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 1992.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal Bayesian deep learning. *Transactions on Machine Learning Research*, 2024.

Gert de Cooman and Jasper De Bock. Randomness is inherently imprecise. *International Journal of Approximate Reasoning*, 2022.

Rabanus Derr and Robert C Williamson. Forecast evaluation and the relationship of regret and calibration. *arXiv preprint arXiv:2401.14483*, 2025.

Rabanus Derr, Jessie Finocchiaro, and Robert C Williamson. Three types of calibration with properties and their semantic and formal relationships. *arXiv preprint arXiv:2504.18395*, 2025.

Daniel Ellsberg. *Risk, Ambiguity, and Decision*. PhD thesis, Harvard University, Cambridge, MA, 1962. Original thesis submitted to the Harvard Economics department. A groundbreaking work in Decision Theory.

Dean P Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 2018.

Christian Fröhlich and Robert C Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 1989.

Parikshit Gopalan and Lunjia Hu. Calibration through the lens of indistinguishability. *SIGecom Exchanges*, 2025.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

Igor I. Gorban. *The Statistical Stability Phenomenon*. Springer Cham, 2017.

Peter Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 2018.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 2022.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *International Conference on Machine Learning*, 2018.

Peter Hoff. Admissibility and complete classes. Lecture notes, 2013. URL https://www2.stat.duke.edu/~pdh10/Teaching/581/LectureNotes/admiss.pdf. Statistics 581 course notes.

Jessica Hullman. Calibration resolves epistemic uncertainty by giving predictions that are indistinguishable from the true probabilities—why is this still unsatisfying? Statistical Modeling, Causal Inference, and Social Science blog, 2024. Accessed: 2025-10-20.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. In *Workshop on Computational Learning Theory*, 1992.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization, 2025.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *International Conference on Learning Representations*, 2025.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, 2020.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Juan Carlos Perdomo. The relative value of prediction in algorithmic decision making. *International Conference on Machine Learning*, 2024.

Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 2022.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 2023.

Stuart J. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. Reliable decisions with threshold calibration. In *Advances in Neural Information Processing Systems*, 2021.

Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, 1954.

Mark J Schervish. A general method for comparing probability assessors. *The annals of statistics*, 1989.

Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 1985.

Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. *International Conference on Machine Learning*, 2024.

Anurag Singh, Siu Lun Chau, and Krikamol Muandet. Truthful elicitation of imprecise forecasts. In *Conference on Uncertainty in Artificial Intelligence*, 2025.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A Naesseth. On continuous monitoring of risk violations under unknown shift. *Conference on Uncertainty in Artificial Intelligence*, 2025.

Rhema Vaithianathan, Diana Benavides Prado, Erin Dalton, and Emily Putnam-Hornstein. Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine*, 2021.

Rajeev Verma. What are good forecasts. Blog post, 2024. Accessed: 2025-11-21.

Rajeev Verma, Volker Fischer, and Eric Nalisnick. On calibration in multi-distribution learning. In *ACM Conference on Fairness, Accountability, and Transparency*, 2025.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press., 1944.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 2021.

A Multi-Distribution Learning: Proofs and Additional Results

A.1 Proof of Proposition 3.1

Proof. First, note that,

$$\{(c(\mathrm{y}, \mathrm{BR}(h(\mathbf{x})))_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \colon h\colon \mathcal{X}\to\Delta\},$$

is finite. This is true by the finiteness of $\mathcal{Y}$ and $\mathcal{A}$. It follows that,

$$\mathrm{co}\{(c(\mathrm{y}, \mathrm{BR}(h(\mathbf{x})))_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \colon h\colon \mathcal{X}\to\Delta\},$$

is closed. Hence, we can apply Theorem 5.2 in [Grünwald and Dawid, 2004], which guarantees the existence of,

$$h_c \in \operatorname*{arg\,min}_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[[c(\mathrm{y}, \mathrm{BR}(h(\mathbf{x})))].$$

The inequality follows immediately by the definition of h_c. □

A.2 Additional Results

Proposition A.1. *For every strictly proper loss function $\ell\colon \mathcal{Y}\times\Delta\to[0,1]$ such that $\{(\ell(\mathrm{y}, h(\mathbf{x}))_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \colon h\colon \mathcal{X}\to\Delta\}$ is closed, there exists a decision maker with dis-utility function $c\colon \mathcal{Y}\times\mathcal{A}\to[0,1]$, finite action set $\mathcal{A}$, and best response mechanism* BR*, and a closed, convex set of distributions $\mathcal{Q}$ on $\mathcal{X}\times\mathcal{Y}$, such that for a perfect predictor h_ℓ (Lemma A.2) it holds, for some $\epsilon>0$,*

$$\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] \geq \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))] + \epsilon.$$

where $h_c\colon \mathcal{X}\to\Delta$, such that,

$$\inf_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] \geq \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))].$$

Proof. Define the set $\mathcal{Q}_\ell \in \arg\max_{Q\in\Delta^{|\mathcal{X}\times\mathcal{Y}|}} \inf_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$. By Theorem 5.1 in [Grünwald and Dawid, 2004] $\mathcal{Q}_\ell \neq \emptyset$. Furthermore, note that,

$$\begin{aligned}
\max_{Q\in\Delta^{|\mathcal{X}\times\mathcal{Y}|}} \inf_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))] &= \max_{Q\in\Delta^{|\mathcal{X}\times\mathcal{Y}|}} \inf_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}[\mathbb{E}_{\mathrm{y}\sim Q_{\mathrm{y}|\mathbf{x}}}[\ell(\mathrm{y}, h(\mathbf{x}))]] \\
&= \max_{Q\in\Delta^{|\mathcal{X}\times\mathcal{Y}|}} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\inf_{P\in\Delta} \mathbb{E}_{\mathrm{y}\sim Q_{\mathrm{y}|\mathbf{x}}}[\ell(\mathrm{y}, P)]\right] \\
&= \max_{Q\in\Delta^{|\mathcal{X}\times\mathcal{Y}|}} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}[\mathbb{E}_{\mathrm{y}\sim Q_{\mathrm{y}|\mathbf{x}}}[\ell(\mathrm{y}, Q_{\mathrm{y}|\mathbf{x}})]] \\
&= \max_{Q_{\mathbf{x}}\in\Delta^{|\mathcal{X}|}} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\max_{Q_{\mathrm{y}|\mathbf{x}}\in\Delta} \mathbb{E}_{\mathrm{y}\sim Q_{\mathrm{y}|\mathbf{x}}}[\ell(\mathrm{y}, Q_{\mathrm{y}|\mathbf{x}})]\right] \\
&= \max_{Q_{\mathbf{x}}\in\Delta^{|\mathcal{X}|}} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}[\mathbb{E}_{\mathrm{y}\sim P_\ell}[\ell(\mathrm{y}, P_\ell)]] \\
&= \mathbb{E}_{\mathrm{y}\sim P_\ell}[\ell(\mathrm{y}, P_\ell)]],
\end{aligned}$$

by strict propriety of ℓ and

$$P_\ell := \max_{P\in\Delta} \mathbb{E}_{\mathrm{y}\sim P}[\ell(\mathrm{y}, P)].$$

Hence, for every $Q_\ell \in \mathcal{Q}_\ell$, $Q_\ell(\cdot|\mathbf{x}) = P_\ell$.

Let $c\colon \mathcal{Y}\times\mathcal{A}\to[0,1]$ be such that,

$$P_\ell \notin \mathcal{P}_c := \operatorname*{arg\,max}_{P\in\Delta} \mathbb{E}_{\mathrm{y}\sim P}[c(\mathrm{y}, \mathrm{BR}(P))].$$

Such a c exists, by the following construction: Pick $y^*\in\mathcal{Y}$ such that $P_\ell(\mathrm{y}=y^*) \leq P_\ell(\mathrm{y}=y')$ for all $y'\in\mathcal{Y}$. Define, with action set $\mathcal{A}=\{A,B\}$,

$$c(\mathrm{y}, a) := \llbracket y=y^*, a=A\rrbracket + (P_\ell(\mathrm{y}=y^*)+\epsilon)\llbracket a=B\rrbracket.$$

where $\epsilon > 0$ such that $P_\ell(\mathrm{y} = y^*) + \epsilon \leq 1$. Then, up to tie-breaking – which can be ignored because of the specific choice of P_c later –,

$$\begin{aligned}
\mathrm{BR}(P) &= \operatorname*{arg\,min}_{a\in\mathcal{A}} \mathbb{E}_{\mathrm{y}\sim P}[c(\mathrm{y}, a)] \\
&= \operatorname*{arg\,min}_{a\in\mathcal{A}} P(\mathrm{y} = y^*)[\![a = A]\!] + (P_\ell(\mathrm{y} = y^*) + \epsilon)[\![a = B]\!] \\
&= \begin{cases} A & \text{if } P(\mathrm{y} = y^*) < (P_\ell(\mathrm{y} = y^*) + \epsilon) \\ B & \text{otherwise.} \end{cases}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\max_{P\in\Delta} \mathbb{E}_{\mathrm{y}\sim P}[c(\mathrm{y}, \mathrm{BR}(P))] &= \max_{P\in\Delta} \mathbb{E}_{\mathrm{y}\sim P}[[\![y = y^*, \mathrm{BR}(P) = A]\!] + (P_\ell(\mathrm{y} = y^*) + \epsilon)[\![\mathrm{BR}(P) = B]\!]] \\
&= \max_{P\in\Delta} P(\mathrm{y} = y^*)[\![P(\mathrm{y} = y^*) < (P_\ell(\mathrm{y} = y^*) + \epsilon)]\!] \\
&+ (P_\ell(\mathrm{y} = y^*) + \epsilon)[\![P(\mathrm{y} = y^*) \geq (P_\ell(\mathrm{y} = y^*) + \epsilon)]\!]] \\
&= P_\ell(\mathrm{y} = y^*) + \epsilon.
\end{aligned}$$

Hence, for all $P \in \mathcal{P}_c$, $P(\mathrm{y} = y^*) \geq P_\ell(\mathrm{y} = y^*) + \epsilon$. Clearly, $P_\ell \notin \mathcal{P}_c$.

Define the set $\mathcal{Q}$ such that it consists of the convex closure of two distributions Q_1 and Q_2 with $Q_1(\cdot|\mathbf{x}) = P_\ell$ respectively $Q_2(\cdot|\mathbf{x}) = P_c$ for some $P_c \in \mathcal{P}_c$.

By construction, it follows that, $h_\ell = P_\ell$ for,

$$h_\ell \in \operatorname*{arg\,min}_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))],$$

respectively $h_c = P_c$ for,

$$h_c \in \operatorname*{arg\,min}_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h(\mathbf{x})))].$$

Because,

$$\begin{aligned}
\mathbb{E}_{\mathrm{y}\sim P_c}[c(\mathrm{y}, \mathrm{BR}(P_c))] &= \mathbb{E}_{\mathrm{y}\sim P_c}[c(\mathrm{y}, B)] \\
&= P_\ell(\mathrm{y} = y^*) + \epsilon \\
&\leq P_c(\mathrm{y} = y^*) \\
&= \mathbb{E}_{\mathrm{y}\sim P_c}[c(\mathrm{y}, A)] \\
&= \mathbb{E}_{\mathrm{y}\sim P_c}[c(\mathrm{y}, \mathrm{BR}(P_\ell))],
\end{aligned}$$

it follows, for all $Q_c \in \arg\max_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})],$

$$\mathbb{E}_{Q_c}[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))] \leq \mathbb{E}_{Q_c}[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] - \epsilon.$$

Hence, for all $Q_c \in \mathcal{Q}_c$,

$$\begin{aligned}
&\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))] \\
&= \inf_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h(\mathrm{v}\mathbf{x})))] \\
&= \inf_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_{Q_c}[c(\mathrm{y}, \mathrm{BR}(h(\mathrm{v}\mathbf{x})))] \\
&= \mathbb{E}_{Q_c}[c(\mathrm{y}, \mathrm{BR}(h_c(\mathbf{x})))] \\
&\leq \mathbb{E}_{Q_c}[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] - \epsilon \\
&\leq \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] - \epsilon.
\end{aligned}$$

□

Lemma A.2. (Existence of perfect MDL predictor). *For a loss function $\ell\colon \mathcal{Y}\times\Delta \to [0,1]$ such that $\{(\ell(\mathrm{y}, h(\mathbf{x}))_{(\boldsymbol{x},y)\in\mathcal{X}\times\mathcal{Y}} \colon h\colon \mathcal{X}\to\Delta\}$ is closed,*[5] *there exists a* perfect predictor, *$h_\ell = \arg\min_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$. Furthermore, if ℓ is strictly proper, then $h_\ell = Q_{\mathrm{y}|\mathbf{x}}$, for some $Q \in \arg\max_{Q\in\mathcal{Q}} \inf_{h\colon \mathbf{x}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$*[6].

[5] Note that this condition is fulfilled for instance if ℓ takes on finitely many different values.

[6] In this statement, we take $\mathcal{Q}$ to be a closed convex set of distributions on $\mathcal{X}\times\mathcal{Y}$. The closure is taken with respect to the Euclidean topology induced on the finite simplex $\Delta(\mathcal{X}\times\mathcal{Y})$

Proof. 1. All conditions for Theorem 5.2 in [Grünwald and Dawid, 2004] are fulfilled, which guarantees the existence of h_ℓ as defined above.

2. By Theorem 5.2 and Theorem 4.1 in [Grünwald and Dawid, 2004],

$$h_\ell \in \operatorname*{arg\,min}_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))],$$

for some $Q \in \arg\max_{Q\in\mathcal{Q}} \inf_{h\colon \mathcal{X}\to\Delta} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathbf{x}))]$. The claim follows by strict propriety of ℓ.

□

Proposition A.3 (Proposition 3.5 in [Fröhlich and Williamson, 2024])**.** *Let $\mathcal{X}\times\mathcal{Y}$ be a finite input-output set where $\mathcal{Y}=\{0,1\}$. There exists $\ell\colon \mathcal{Y}\times\Delta\to\mathbb{R}$, a convex closed set $\mathcal{Q}$ of distributions on $\mathcal{X}\times\mathcal{Y}$, a ℓ-optimal hypothesis,*

$$h_\ell \in \operatorname*{arg\,min}_{h\colon \mathcal{X}\to\Delta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[\ell(\mathrm{y}, h(\mathrm{v}\mathbf{x}))],$$

such that,

$$\sup_{Q\in\mathcal{Q}} \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{x}}}\left[\mathbb{E}_{\mathrm{y}\sim Q_{\mathrm{y}|\mathbf{x}}}[\ell(\mathrm{y}, h_\ell(\mathbf{x}))] - \mathbb{E}_{\mathrm{y}\sim h_\ell(\mathbf{x}))}[\ell(\mathrm{y}, h_\ell(\mathbf{x}))]\right] > 0.$$

Proof. By Lemma A.2 $h_\ell = Q_{\mathrm{y}|\mathbf{x}}$ for some $Q\in\mathcal{Q}$. Pick a $Q'\in\mathcal{Q}$ such that $Q_{\mathrm{y}}(\mathrm{y}=\tilde{y}) < Q'_{\mathrm{y}}(\mathrm{y}=\tilde{y})$ for some $\tilde{y}\in\mathcal{Y}$ and $Q_{\mathbf{x}} = Q'_{\mathbf{x}}$, which exists by the assumption on differing $\mathcal{Y}$-marginals. Then, define,

$$c(\mathrm{y}, a) := \begin{cases} 0 & \text{if } y \neq \tilde{y} \\ 1 & \text{otherwise .} \end{cases}$$

It follows,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim Q'_{\mathbf{x}}}\left[\mathbb{E}_{\mathrm{y}\sim Q'_{\mathrm{y}|\mathbf{x}}}[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))] - \mathbb{E}_{\mathrm{y}\sim h_\ell(\mathbf{x}))}[c(\mathrm{y}, \mathrm{BR}(h_\ell(\mathbf{x})))]\right] \\
&= \mathbb{E}_{\mathbf{x}\sim Q'_{\mathbf{x}}}\left[Q'_{\mathrm{y}|\mathbf{x}}(\mathrm{y}=\tilde{y}) - Q_{\mathrm{y}|\mathbf{x}}(\mathrm{y}=\tilde{y})]\right] \\
&= \mathbb{E}_{\mathbf{x}\sim Q'_{\mathbf{x}}}\left[Q'_{\mathrm{y}|\mathbf{x}}(\mathrm{y}=\tilde{y})\right] - \mathbb{E}_{\mathbf{x}\sim Q'_{\mathbf{x}}}\left[Q_{\mathrm{y}|\mathbf{x}}(\mathrm{y}=\tilde{y})]\right] \\
&= Q'_{\mathrm{y}}(\mathrm{y}=\tilde{y}) - Q_{\mathrm{y}}(\mathrm{y}=\tilde{y}) \\
&> 0.
\end{aligned}$$

□

B Simultaneous Calibration

B.1 Reformulation and Proof of Proposition 3.2

Proposition B.1. *Let $\mathcal{X}$ be the input space and labels $\mathcal{Y}=\{0,1\}$. Consider two distributions P^1 and P^2 on $(\mathbf{x}, \mathrm{y})$ such that their feature marginals agree, $P^1_{\mathbf{x}} = P^2_{\mathbf{x}}$, but their base rates differ, $\mathbb{E}_{P^1}[\mathrm{y}] \neq \mathbb{E}_{P^2}[\mathrm{y}]$. Then no predictor $h:\mathcal{X}\to[0,1]$ can be calibrated for both P^1 and P^2.*

Proof. Calibration of h for P^i means

$$\mathbb{E}_{P^i}\big[\mathrm{y} \mid h(\mathbf{x}) = p\big] = p \quad \text{a.s.} \quad \forall p$$

Taking expectations and applying the tower property,

$$\mathbb{E}_{P^i}[\mathrm{y}] \;=\; \mathbb{E}_{P^i}\big[\mathbb{E}[\mathrm{y} \mid h(\mathbf{x}) = p]\big] \;=\; \mathbb{E}_{P^i}[h(\mathbf{x})].$$

The quantity $\mathbb{E}_{P^i}[h(\mathbf{x})]$ depends only on the marginal $P^i_{\mathbf{x}}$ (since h is a function of $\mathbf{x}$ alone). Because $P^1_{\mathbf{x}} = P^2_{\mathbf{x}}$, we have $\mathbb{E}_{P^1}[h(\mathbf{x})] = \mathbb{E}_{P^2}[h(\mathbf{x})]$, hence $\mathbb{E}_{P^1}[\mathrm{y}] = \mathbb{E}_{P^2}[\mathrm{y}]$, contradicting the hypothesis.

□

B.2 The $\mathbf{x}$-Marginal Irrelevance and How to Avoid It

If a single predictor $h(\mathbf{x})$ is used across environments that share the same $P_{\mathbf{x}}$, then calibration forces

$$\mathbb{E}_{P^i}[\mathrm{y}] = \mathbb{E}[h(\mathbf{x})] \qquad \text{for all } i,$$

i.e., the mean score is pinned solely by the common $\mathbf{x}$-marginal. Consequently, environments with different base rates cannot share a simultaneously calibrated $h(\mathbf{x})$. If the predictor may depend on the environment index i, i.e., we use $\tilde{h}(\mathbf{x}, i)$ and impose per-environment calibration

$$\mathbb{E}\big[\mathrm{y} \mid \tilde{h}(\mathbf{x}, i) = p,\, i\big] = \tilde{p} \quad \text{a.s.,} \quad \forall p,$$

then $\mathbb{E}_{P^i}[\mathrm{y}] = \mathbb{E}_{P^i}[\tilde{h}(\mathbf{x}, i)]$ can vary with i even when $P^1_{\mathbf{x}} = P^2_{\mathbf{x}}$. Thus the coupling through the shared $\mathbf{x}$-marginal disappears because the predictor itself changes with i. In fairness, for example, this would require using sensitive attribute information to design the predictor. However, it may be prohibited to do so. There could be other features that might be correlated with the sensitive attributes. However, in the absence of any such correlated information, simultaneous calibration can be tricky, and hence could lead to disparate impact.

B.3 Calibration Under Distributional Ambiguity

Lemma B.1 establishes that when two environments P^1 and P^2 share the same feature marginal $P^1_{\mathbf{x}} = P^2_{\mathbf{x}}$ but differ in their label base rates $\mathbb{E}_{P^1}[\mathrm{y}] \neq \mathbb{E}_{P^2}[\mathrm{y}]$, no single predictor $h : \mathcal{X} \to [0, 1]$ can be simultaneously calibrated for both. This impossibility result has an important implication for prediction under uncertainty about the data-generating environment.

If the environment index i were observable, one could define environment-specific predictors $h(\mathbf{x}, i)$ and enforce per-environment calibration,

$$\mathbb{E}[\mathrm{y} \mid h(\mathbf{x}, i) = p,\, i] = p, \qquad \forall i, \quad \forall p.$$

In this case, the coupling through the shared marginal $P_{\mathbf{x}}$ vanishes, and each $h(\cdot, i)$ may correctly reflect the distinct base rate $\mathbb{E}_{P^i}[\mathrm{y}]$.

However, if the predictor cannot depend on i, i.e., if we must produce a single mapping $h(\mathbf{x})$ without knowing which P^i governs test-time data, then exact calibration for all admissible environments becomes impossible. The only coherent alternative, if one wishes to preserve honest uncertainty about predictive accuracy, is to *adopt an imprecise or set-valued notion of prediction.*

Formally, rather than a single score $h(\mathbf{x})$, we consider a set-valued predictor

$$H(\mathbf{x}) \;=\; \big\{\, h_i(\mathbf{x}) : h_i(\mathbf{x}) = \mathbb{E}_{P^i}[\mathrm{y} \mid \mathbf{x}],\; i \in \mathcal{I} \,\big\},$$

where each h_i is calibrated with respect to its corresponding environment P^i. The set $H(\mathbf{x})$ represents the range of calibrated conditional expectations consistent with the possible environments.

Thus, when the environment cannot be conditioned on explicitly, any single precise probability necessarily fails calibration under some P^i. To remain probabilistically truthful, the predictor must acknowledge this ambiguity through imprecision, reporting a set or interval of plausible predictions rather than a single number.

C Testing by Betting of Imprecise Forecasts

We provide further details for the betting-based evaluation procedure introduced in Section 5. Given the interval forecast $[a, b]$, let $\{y_i\}_{i \geq 1} \subset \{0, 1\}$ denote the observed outcomes. To accommodate adaptive selection of subsequences, we introduce a *predictable selector* $A_i \in \{0, 1\}$ that decides whether to bet on round i. The bettor maintains two wealth processes:

$$M^U_t = \prod_{i=1}^{t} \left(1 + A_i \lambda_i (\mathrm{y}_i - b)\right), \qquad M^L_t = \prod_{i=1}^{t} \left(1 + A_i \zeta_i (a - y_i)\right),$$

where λ_i and ζ_i are *betting rates* chosen predictably (possibly as functions of the past) from the safe ranges

$$\lambda_i \in \Big[0,\, \tfrac{1}{b}\Big), \qquad \zeta_i \in \Big[0,\, \tfrac{1}{1-a}\Big].$$

These bounds ensure non-negativity of each factor in the above wealth processes.

Supermartingale validity. Under the *upper-bound null*

$$H_0^U : \quad \mathbb{E}[\mathrm{y}_i \mid \mathcal{F}_{i-1}] \leq b \quad \forall i,$$

we have

$$\mathbb{E}[1 + A_i\lambda_i(\mathrm{y}_i - b) \mid \mathcal{F}_{i-1}] = 1 + A_i\lambda_i(\mathbb{E}[y_i \mid \mathcal{F}_{i-1}] - b) \leq 1.$$

Hence $(M_t^U)_{t\geq 1}$ is a nonnegative supermartingale. Analogously, $(M_t^L)_{t\geq 1}$ is a supermartingale under

$$H_0^L : \quad \mathbb{E}[y_i \mid \mathcal{F}_{i-1}] \geq a \quad \forall i.$$

By Ville's inequality, both tests are *anytime-valid*:

$$\Pr_{H_0^U}\Big(\sup_t M_t^U \geq 1/\alpha\Big) \leq \alpha, \qquad \Pr_{H_0^L}\Big(\sup_t M_t^L \geq 1/\alpha\Big) \leq \alpha.$$

Thus, a large observed wealth constitutes valid evidence against the corresponding null.

Optional skipping and mixtures. The inclusion of the selector A_i enables *optional skipping*: the bettor may restrict bets to predictable subsequences (e.g., even rounds or context-defined events) without affecting validity. Furthermore, power can be increased by forming convex mixtures of several betting strategies or rate grids:

$$\widetilde{M}_t^U = \sum_j w_j M_t^{U,(j)}, \quad w_j \geq 0, \ \sum_j w_j = 1,$$

since mixtures of e-processes remain e-processes.

Interpretation. Under the nulls, expected wealth does not increase, whereas systematic growth of M_t^U or M_t^L provides evidence that the forecast interval is misspecified—either its upper or lower bound is violated. Running both processes in parallel therefore tests *calibration* and *size* simultaneously: intervals that are too narrow or systematically biased are eventually falsified, while correctly calibrated intervals remain within the controlled error budget (α). The test enjoys asymptotic power 1 and is valid under adaptive stopping and adaptive subsequence selection. The methodology is inspired by Waudby-Smith and Ramdas [2024], and was recently applied to detect violations of bounded means in an arbitrary non-*i.i.d.* stream data by Timans et al. [2025].