Proceedings of Machine Learning Research 329:1–23, 2026 Conformal and Probabilistic Prediction with Applications

Venn Predictive Decision-Making

Johan Hallberg Szabadváry JOHAN.HALLBERG.SZABADVARY@JU.SE
Jönköping University and Stockholm University, Sweden

Lars Carlsson LARS.CARLSSON@JU.SE
Jönköping University, Sweden

Editors: Ernst Ahlberg, Ulf Johansson, Henrik Boström, Alberto Carlevaro, Johan Hallberg Szabadváry and Lars Carlsson

Abstract

Venn predictors generate multiprobability predictions, which are sets of probability distributions with provable calibration guarantees, quantifying epistemic ambiguity (Knightian uncertainty) rather than providing a single probability estimate. This paper introduces Venn predictive decision-making systems (Venn-PDMSs), a framework that integrates multiprobability predictions with decision-theoretic principles to guide decision-making under uncertainty. As Venn predictors explicitly account for epistemic ambiguity, the resulting Venn-PDMSs possess the capability to "know what they do not know," thereby facilitating safe human-in-the-loop automation in numerous scenarios. We formalise utility- and regret-based decision criteria, parametrised by an optimism index that allows for interpolation between pessimistic and optimistic subjective stances towards uncertainty. We establish theoretical guarantees, including invariance under affine transformations of the utility function and convergence to Bayes-optimal decisions in certain cases. Through synthetic examples and straightforward applications to mushroom classification and financial credit assessment, we demonstrate that Venn-PDMSs effectively balance epistemic ambiguity and risk (which is probabilistically quantifiable). We also highlight the limitations of aggregating multiprobabilities into point estimates and emphasise the importance of explicitly incorporating ambiguity-aware decision criteria in high-stakes predictive decision-making.

Keywords: Venn prediction, Predictive decision-making, Utility function, Decision criteria

1. Introduction

Venn predictors were introduced in Vovk (2003) and described in detail in Vovk et al. (2022, Chapter 6). They are *multiprobability predictors*; instead of announcing a single probability distribution over the label space, they are allowed to propose several probability distributions as their prediction. The resulting *multiprobability prediction* is considered valid if any of its components is valid. If all components are close to each other, the multiprobability prediction provides a useful probabilistic prediction; if not, we are essentially in a situation of Knightian uncertainty, and the diverse multiprobability prediction serves as a useful warning. This property of Venn predictors, namely, their "knowing what they do not know" coupled with their guaranteed validity, makes them suitable for application in decision-making under uncertainty, particularly when the stakes are high.

Vovk and Bendtsen (2018) integrated *conformal predictive systems* (CPSs) with the classical von Neumann-Morgenstern utility theory (Von Neumann and Morgenstern, 1947). The resulting *conformal predictive decision-making system* (CPDMS) optimises expected

utility across available decisions for an unknown outcome, using a conformal predictive distribution function (CPD) that is provably well-calibrated. However, since CPSs achieve their calibration through random smoothing, the CPDMS collapses the Knightian uncertainty to a single CPD, thus basing its decisions solely on risk (which, according to Knight (1921), is quantifiable) while neglecting the uncertainty.

In this study, we demonstrate the application of Venn predictors in decision-making. We propose a comprehensive framework for *Venn Predictive Decision-Making Systems* (Venn-PDMS), which couples multiprobability predictions with decision-theoretic principles to guide decision-making under uncertainty. We argue that naively aggregating multiprobabilities into a single probability estimate discards critical information about epistemic ambiguity (Knightian uncertainty) and can lead to suboptimal or misleading decisions. Instead, decision-makers must explicitly incorporate their tolerance for ambiguity through subjective criteria by interpolating between pessimistic and optimistic stances. We formalise this approach using von Neumann–Morgenstern utility theory extended to multiprobability predictions, introducing both utility-based and regret-based decision criteria parameterised by an optimism index.

Our framework provides theoretical guarantees, including invariance under affine transformations of the utility function and asymptotic convergence to Bayes-optimal decisions as epistemic ambiguity vanishes. Through synthetic examples and real-world applications, we demonstrate that Venn-PDMSs effectively balance epistemic ambiguity and risk, outperforming existing conformal predictive decision-making methods in a simple example. These results underscore the limitations of aggregating multiprobabilities into point estimates and highlight the importance of explicitly incorporating ambiguity-aware criteria into high-stakes predictive decision-making processes.

2. Venn Predictors

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two measurable spaces, called the *object space* and *label space*, respectively. We typically refer to their Cartesian product $\boldsymbol{Z} := \boldsymbol{X} \times \boldsymbol{Y}$ as the *example space*, where each example $z = (x, y)$ consists of an object x and its label y. The set of all *bags* (or multisets) of size n of examples $z \in \boldsymbol{Z}$ is denoted $Z^{(n)}$. A bag is similar to a set in that the elements are unordered; however, we allow for the possibility that examples may be repeated. The set $Z^{(n)}$ is measurable and can be defined as the power space $\boldsymbol{Z}^n$ with a non-standard σ-algebra (see Vovk et al. (2022) for details). The union of all $\boldsymbol{Z}^{(n)}$, $n = 0, 1, 2, \ldots$ is denoted $\boldsymbol{Z}^{(*)}$.

A *Venn taxonomy* is a measurable function $K : \boldsymbol{Z}^{(*)} \times \boldsymbol{Z} \to \boldsymbol{K}$ where $\boldsymbol{K}$ is a countable set equipped with the discrete σ-algebra. Elements $\kappa \in \boldsymbol{K}$ are referred to as *categories*. Typically, the category $\kappa_i := K(\{z_1, \ldots, z_n\}, z_i)$ of an example z_i is a kind of classification of z_i, which may depend on the bag $\{z_1, \ldots, z_n\}$. Every Venn taxonomy determines a Venn predictor in the following way. Let $z_1, \ldots, z_n$ be a training sequence of examples, and x_{n+1} be a test object whose label is unknown but could potentially be $y \in \boldsymbol{Y}$. Tentatively, we write $z_{n+1} = (x_{n+1}, y)$. The empirical probability distribution of the labels in the category

containing the test example is[1]

$$P^y\{y'\} := P^y(\{y'\}) := \frac{|\{i = 1, \dots, n+1 : \kappa_i^y = \kappa_{n+1}^y \wedge y_i = y'\}|}{|\{i = 1, \dots, n+1 : \kappa_i^y = \kappa_{n+1}^y\}|} \tag{1}$$

where,

$$\kappa_i^y := K(\lbag z_1, \dots, z_{n+1} \rbag, z_i).$$

P^y is a probability distribution on $\boldsymbol{Y}$. Then Venn predictor determined by the taxonomy K is the multiprobability predictor

$$P := (P^y : y \in \boldsymbol{Y}).$$

It is a family consisting of between one and $|\boldsymbol{Y}|$ distinct probability distributions on $\boldsymbol{Y}$. We refer to P as the *multiprobability prediction*. This note restricts the attention to the binary case when $|\boldsymbol{Y}| = 2$, but most results generalise to larger finite label spaces.

In summary, the multiprobability prediction is constructed by, for every potential label $y \in \boldsymbol{Y}$ of the test object x_{n+1}, computing the empirical probability distribution of the labels that belong to the same category (according to the taxonomy K) as the tentative example (x_{n+1}, y).

2.1. Validity of Venn Predictors

All Venn predictors share the virtue of being valid in various ways. Here (as in Vovk et al. (2022)), we are content with the simplest notion of validity. We say that a random variable P taking values in $\boldsymbol{P}(\boldsymbol{Y})$ (the set of probability distributions on $\boldsymbol{Y}$) is *perfectly calibrated* for a random variable Y taking values in $\boldsymbol{Y}$ if, for every $y \in \boldsymbol{Y}$,

$$\mathbb{P}(Y = y \mid P) = P\{y\} \quad \text{a.s.} \tag{2}$$

We may think of P as the prediction made by some probabilistic predictor for Y. Perfect calibration means that the predictor gets the probabilities right on average. A probabilistic predictor that approximately satisfies (2) is informally said to be well-calibrated.

Let Y be the measurable function $Y : \boldsymbol{Z}^{n+1} \to \boldsymbol{Y}$ that maps $(z_1, \dots, z_{n+1})$ to y_{n+1}, the true label of the test example z_{n+1}. This is what we want to predict. A *selector* is a measurable function $S : \boldsymbol{Z}^{n+1} \to \boldsymbol{Y}$. Both Y and S are $\boldsymbol{Y}$-valued random variables. The probability measure P^y in (1) is a measurable function of $(z_1, \dots, z_{n+1})$ (which does not depend on the label of z_{n+1}), so it is a $\boldsymbol{P}(\boldsymbol{Y})$-valued random variable (for a fixed $y \in \boldsymbol{Y}$). The composition P^S, obtained by plugging S in place of y, is also a $\boldsymbol{P}(\boldsymbol{Y})$-valued random variable.

Proposition 1 (Proposition 6.1 (Vovk et al., 2022)) *There exists a selector S such that, for any Venn predictor, the composition P^S is perfectly calibrated for Y under any power distribution.*

Proposition 1 is a special case of Vovk et al. (2022, Theorem. 12.1). The selector, as one might expect, is $S := Y$. Intuitively, at least one of the at most $|\boldsymbol{Y}|$ distinct probability measures output by a Venn predictor is perfectly calibrated.

1. The symbol "$\wedge$" denotes the logical "and" operator.

3. Predictive Decision-Making Systems

For each object $x \in \boldsymbol{X}$, label $y \in \boldsymbol{Y}$, and decision d chosen from a *decision space* $\boldsymbol{D}$ (not assumed measurable in any way), we associate a *von Neumann-Morgenstern utility* $U(x, y, d)$. Vovk and Bendtsen (2018) remark that, in applications, U typically does not depend on x. However, there are settings, for example, in medical diagnoses, where the utility of prescribing an aggressive treatment given a true disease state ($y = 1$) may be vastly different for a young, otherwise healthy patient than for an elderly patient with severe comorbidities. Formally encoding x into the utility function ensures that the decision-making system accurately models these instance-specific risk profiles.

The problem is that the utility $U(x, y, d)$ depends on the unknown label y of x_n. Vovk and Bendtsen (2018) defined a *predictive decision-making system* (PDMS) as a measurable function $F : \boldsymbol{Z}^* \times \boldsymbol{X} \to \boldsymbol{D}$. The idea is that $F(z_1, \ldots, z_{n-1}, x_n)$ is the decision recommended for the test object x_n. The PDMSs in Vovk and Bendtsen (2018) are randomised, which means that they additionally depend on a random number $\tau \in [0, 1]$. In contrast, this study does not consider randomised PDMSs.

The *regret* of a PDMS F for an object x_n and a training set $z_1, \ldots, z_{n-1}$ under a probability measure Q on $\boldsymbol{Z}$ is defined as

$$\begin{aligned} R_F(z_1, \ldots, z_{n-1}, x_n) := \sup_{d \in \boldsymbol{D}} &\int U(x_n, y, d) Q(dy \mid x_n) \\ &- \int U(x_n, y, F(z_1, \ldots, z_{n-1}, x_n)) Q(dy \mid x_n) \end{aligned}$$

where $Q(dy \mid x)$ is a regular conditional distribution for y given x under Q. Under the standard assumption that $\boldsymbol{Z}$ is a standard Borel space, such regular conditional distributions are guaranteed to exist.

Given a predictive distribution $\Pi(dy \mid z_1, \ldots, z_{n-1}, x_n)$ on $\boldsymbol{Y}$ for the unknown label y, an obvious PDMS selects an ϵ-optimal decision for some strictly positive tolerance $\epsilon > 0$. That is, the recommended decision $F(z_1, \ldots, z_{n-1}, x_n) \in \boldsymbol{D}$ is chosen such that

$$\int_{\boldsymbol{Y}} U(x_n, y, F(z_1, \ldots, z_{n-1}, x_n)) \Pi(dy \mid x_n) \geq \sup_{d \in \boldsymbol{D}} \int_{\boldsymbol{Y}} U(x_n, y, d) \Pi(dy \mid x_n) - \epsilon$$

where $\Pi(\cdot \mid x_n)$ is a shorthand for the predictive distribution conditioned on the training set and test object.

Remark 2 *In standard machine learning settings, the supremum is often guaranteed to be attained, such as when the decision space $\boldsymbol{D}$ is finite (the assumption made by Vovk and Bendtsen (2018)) or when $\boldsymbol{D}$ is a compact topological space and the utility function $U(x, y, \cdot)$ is upper semi-continuous; in this case, we can set $\epsilon = 0$. The PDMS then simplifies to the standard expected utility maximiser*

$$F(z_1, \ldots, z_{n-1}, x_n) \in \operatorname*{arg\,max}_{d \in \boldsymbol{D}} \int_{\boldsymbol{Y}} U(x_n, y, d) \Pi(dy \mid x_n).$$

Remark 3 *In statistical decision theory, it is customary to base decisions on a loss function $L(x, y, d) = -U(x, y, d)$. Berger (1993) jokingly remarked that "Statisticians seem to be*

pessimistic creatures who think in terms of losses". Sometimes, it can be easier to elicit or describe losses than utilities; in this case, a simple "sign-flip" translates the problem into our utility-maximization framework. More generally, von Neumann–Morgenstern utility functions are unique only up to a positive affine transformation. That is, replacing $U(x, y, d)$ *with* $aU(x, y, d) + b$ *for any constants* $a > 0$ *and* $b \in \mathbb{R}$ *will not alter the ordering of preferences or the resulting optimal decisions. This affords us the practical convenience of arbitrarily shifting the baseline (e.g., setting the worst possible outcome to zero) or scaling the units (e.g., translating between dollars and abstract utility points) without affecting the behaviour of the PDMS.*

4. Venn Predictive Decision-Making Systems

We now turn to the problem of how to define PDMSs based on Venn predictors, or more generally, multiprobability predictors. The problem, of course, is that multiprobability predictions do not define a single predictive probability distribution to integrate. Throughout the rest of this note, we assume (if not stated otherwise) that $\boldsymbol{Y} = \{0, 1\}$. Thus, any probability distribution Q on $\boldsymbol{Y}$ is represented by a single number $p := Q\{1\}$, and a multiprobability prediction can be written $P = (p^y : y \in \{0, 1\}) = (p^0, p^1)$. We have no theoretical justification to assign a probability distribution over p^0 and p^1 or to aggregate them into a single number. Rather, we view the multiprobabilistic prediction as representing two possible states of Reality, and we have no idea which state it actually chooses.

In this binary setting, we can simplify the notation for the expected utility. As we have two valid states, any decision $d \in D$ yields two distinct, expected utilities. We define the state-specific expected utility E_i for $i \in \{0, 1\}$ as

$$E_i(x, d) = p^i U(x, 1, d) + (1 - p^i) U(x, 0, d) \tag{3}$$

We also define the state-specific expected regret $R_i(x, d)$, which represents the expected opportunity cost incurred if state i is true, as follows:

$$R_i(x, d) = \left(\max_{d' \in D} E_i(x, d') \right) - E_i(x, d)$$

Because a Venn PDMS must evaluate the utility of an action across a set of distributions rather than a single distribution, the standard arg max formulation is insufficient. Instead, a PDMS must employ an ambiguity-averse decision criterion that reflects a subjective preference under uncertainty (optimism or pessimism) to resolve the multiprobability prediction (p^0, p^1).

Following Knight (1921) and the observations of Vovk et al. (2022, Section 5.6.1), we distinguish between risk and uncertainty. Pure risk is probabilistically quantifiable, whereas pure uncertainty is not. By defining the width of the multiprobability $\Delta p := |p^1 - p^0|$, we are not assigning a higher-order probability distribution to the uncertainty, which would impermissibly collapse it back into risk. Rather, Δp quantifies the magnitude of our ignorance; if it is small, we can be rather certain that both p^0 and p^1 are good estimates (one of them is perfectly calibrated and the other is very close). In contrast, if Δp is large, we know that we have a good estimate, but not which of the two numbers represents it. The

uncertainty remains strictly Knightian: there is no justification for assigning probabilities to p^0 and p^1.

Remark 4 (Terminology: Uncertainty vs. Ambiguity) *Knight (1921) famously distinguished between measurable* risk *(where probabilities are known) and unmeasurable* uncertainty *(where probabilities are unknown). However, because the term "uncertainty" has become heavily overloaded in modern machine learning, often conflated with softmax entropy, predictive variance, or out-of-distribution detection, we adopt the term* epistemic ambiguity *(Ellsberg, 1961) to refer to Knight's strict uncertainty. This deliberate choice resolves semantic ambiguity and formally aligns our framework with the economic literature on ambiguity aversion.*

To evaluate actions without assuming whether p^0 or p^1 is the "true" state, we define the strict lower and upper bounds for both utility and regret:

$$\begin{aligned}
\text{Lower Expected Utility (Pessimistic):} \quad & \underline{E}(x,d) = \min\{E_0(x,d), E_1(x,d)\}. \\
\text{Upper Expected Utility (Optimistic):} \quad & \overline{E}(x,d) = \max\{E_0(x,d), E_1(x,d)\}. \\
\text{Lower Expected Regret (Optimistic):} \quad & \underline{R}(x,d) = \min\{R_0(x,d), R_1(x,d)\}. \\
\text{Upper Expected Regret (Pessimistic):} \quad & \overline{R}(x,d) = \max\{R_0(x,d), R_1(x,d)\}.
\end{aligned}$$

4.1. Decision Criteria under Ambiguity

A Venn PDMS maps the index-agnostic bounds—the lower and upper expected utilities and regrets defined above, which do not presuppose which of p^0 or p^1 is the correct distribution—to a recommended decision $d \in \boldsymbol{D}$ using classical decision criteria. Because the true state of Reality is unknown, the system's behaviour, given a multiprobability prediction, is dictated entirely by subjective preference (optimism vs. pessimism) encoded in the chosen criterion.

The foundational literature on decision-making under strict ignorance (see, e.g., Luce and Raiffa (1957) for a comprehensive critical survey) provides several competing philosophical stances to resolve this ambiguity. Notably, Wald (1950) advocated for a strictly pessimistic approach that evaluates an action solely by its worst-case utility guarantee, ensuring absolute baseline survival. In contrast, Savage (1951) argued that rational decision-makers should instead guard against maximum opportunity cost, formally introducing the minimax regret criterion. Building on these established philosophies, we consider four "pure", extreme stances towards epistemic ambiguity

Pure Criteria The four extreme philosophical stances towards epistemic ambiguity are

- **Maximin Utility (Wald):** Maximizes the worst-case utility, $\underline{E}(x,d)$.
- **Maximax Utility:** Maximizes the best-case utility, $\overline{E}(x,d)$.
- **Minimax Regret (Savage):** Minimizes the worst-case opportunity cost, $\overline{R}(x,d)$.
- **Minimin Regret:** Minimizes the best-case regret, $\underline{R}(x,d)$.

Remark 5 (Utility, Regret, and the Axiom of Irrelevant Alternatives) *The consideration of regret-based decision criteria under epistemic ambiguity highlights a classical*

tension in decision theory. Savage's minimax regret (Savage, 1951) strictly violates the von Neumann–Morgenstern axiom of independence of irrelevant alternatives *(IIA). As pointed out by Chernoff (1954), IIA dictates that a rational agent's preference between two decisions should not change if a third, unchosen option is added to the decision space. Because Savage's minimax regret evaluates a decision relative to the maximum possible utility across all available decisions in a given state, it is inherently "menu-dependent." Introducing a new option can, at least theoretically, flip preferences between existing choices.*

Savage originally proposed the minimax regret strategy as a solution to the paralysing conservatism of Wald's maximin utility rule (Wald, 1950). By strictly optimising the worst possible utility, Wald's criterion in effect assumes that Reality is acting adversarially. Regret softens this pessimism by considering the relative opportunity cost rather than the absolute worst-case performance. Interestingly, Savage later abandoned his regret criterion in his seminal work Savage (1954), formulating subjective expected utility and arguing that rational agents must always form exact subjective point probabilities. However, as shown in Vovk et al. (2022, Chapter 5), probabilistic prediction in the sense of estimating true probabilities from a given finite training set is essentially impossible under the assumption of unconstrained randomness. This is precisely why Venn predictors output a multiprobability prediction rather than a probability distribution. For this reason, modern robust decision theory (Gilboa and Schmeidler, 1989; Manski, 2004) has resurrected Savage's minimax regret as a pragmatic ambiguity-aware mechanism that navigates finite-sample ambiguity without necessitating the extreme pessimism of Wald (1950).

Interpolation between optimism and pessimism. To formalise a specific, tunable subjective stance towards epistemic ambiguity, we interpolate between the pessimistic and optimistic bounds using an optimism index $\alpha \in [0, 1]$ (Hurwicz, 1951). We construct two parallel tracks for decision-making

1. **Utility-Based Interpolation:** Evaluates actions based on absolute expected utility by maximizing the functional
$$H_\alpha^{\mathrm{U}}(x, d) = \alpha\overline{E}(x, d) + (1 - \alpha)\underline{E}(x, d).$$
Setting $\alpha = 0$ recovers Wald's criterion, whereas $\alpha = 1$ recovers maximax.

2. **Regret-Based Interpolation:** Evaluates actions based on expected opportunity cost. Because regret is a penalty, an optimistic decision-maker expects the lowest possible regret, while a pessimistic one guards against the highest possible regret. We seek to minimize the functional
$$H_\alpha^{\mathrm{R}}(x, d) = \alpha\underline{R}(x, d) + (1 - \alpha)\overline{R}(x, d).$$
Setting $\alpha = 0$ heavily weights the upper bound, recovering Savage's Minimax Regret, while $\alpha = 1$ recovers Minimin Regret.

Example: Non-Equivalence of Decision Criteria. To illustrate that corner strategies (Wald, maximax, Savage, and minimin regret) represent fundamentally distinct evaluations of ambiguity, we constructed a synthetic decision-making scenario.

We consider a test object x_n with two possible true labels ($y \in \{0, 1\}$) and four available decisions ($d \in \{d_1, d_2, d_3, d_4\}$). We define the utility function $U(x_n, y, d)$ per decision by

- d_1: $U(x_n, 0, d_1) = 100$, $U(x_n, 1, d_1) = 0$
- d_2: $U(x_n, 0, d_2) = 0$, $U(x_n, 1, d_2) = 90$
- d_3: $U(x_n, 0, d_3) = 50$, $U(x_n, 1, d_3) = 50$
- d_4: $U(x_n, 0, d_4) = 60$, $U(x_n, 1, d_4) = 40$

Assume that the predictor outputs the multiprobability prediction $P = (p^0, p^1) = (0.1, 0.9)$. We calculate the state-specific expected utilities using (3). Table 1 lists the resulting expected utility bounds and corresponding state-specific regrets for each decision. The maximum possible expected utility is 90 if state 0 is true and 81 if state 1 is true.

	Expected Utility Bounds				**Regret Bounds**			
Action (d)	E_0	E_1	$\underline{E}$	$\overline{E}$	R_0	R_1	$\overline{R}$	$\underline{R}$
d_1	90	10	10	**90**	0	71	71	**0**
d_2	9	81	9	81	81	0	81	**0**
d_3	50	50	**50**	50	40	31	40	31
d_4	58	42	42	58	32	39	**39**	32

Table 1: A synthetic decision matrix demonstrating the mathematical divergence of pure decision criteria under large epistemic ambiguity ($p^0 = 0.1, p^1 = 0.9$, $\Delta p = 0.8$). Bold values indicate the optimal bound for each respective strategy.

Evaluating the four pure strategies on this exact matrix yields completely divergent optimal decisions:

- **Maximin Utility (Wald) ($\alpha = 0$) chooses d_3:** Wald strictly maximizes the lower utility bound ($\underline{E}$). It selects d_3 to guarantee a baseline utility of 50, rejecting d_4 whose expected utility could fall to 42.
- **Maximax Utility ($\alpha = 1$) chooses d_1:** Maximax strictly maximizes the upper utility bound ($\overline{E}$). It greedily selects d_1 for its potential expected reward of 90.
- **Minimax Regret (Savage) ($\alpha = 0$) chooses d_4:** Savage minimizes the upper regret bound. It rejects the extreme safety of d_3 (which could incur an opportunity cost of 40) and selects d_4, strictly mathematically bounding the maximum possible opportunity cost to 39.
- **Minimin Regret ($\alpha = 1$) ties between d_1 and d_2:** Minimin Regret minimizes the lower regret bound ($\underline{R}$). Because d_1 happens to be optimal under State 0 and d_2 happens to be optimal under State 1, both achieve a best-case regret of 0. The functional fails to break the tie, ignoring the absolute utility differences. In such cases, since either decision is equally attractive, ties can be resolved arbitrarily, e.g. via coin tossing.

Remark 6 *In standard decision theory under risk, maximising expected utility is strictly equivalent to minimising expected regret. However, as we have seen, under the epistemic ambiguity of a multiprobability prediction, this equivalence is broken and holds only when the optimism index is exactly* $\alpha = 1/2$*. By definitions of the upper and lower bounds, evaluating the utility criterion at* $\alpha = 1/2$ *yields*

$$H^U_{1/2}(x,d) = \frac{1}{2}\left(\max\{E_0, E_1\} + \min\{E_0, E_1\}\right).$$

Using the identity $\max(A,B) + \min(A,B) = A + B$*, this simplifies to*

$$H^U_{1/2}(x,d) = \frac{1}{2}\left(E_0(x,d) + E_1(x,d)\right).$$

Similarly, for regret, let $M_i = \max_{d' \in D} E_i(x,d')$ *be the maximum attainable utility in state* i*, such that* $R_i(x,d) = M_i - E_i(x,d)$*. Applying the same identity gives*

$$H^R_{1/2}(x,d) = \frac{1}{2}\left(R_0(x,d) + R_1(x,d)\right) = \frac{M_0 + M_1}{2} - H^U_{1/2}(x,d).$$

Because the term $\frac{1}{2}(M_0 + M_1)$ *is a constant independent of the decision* d*, minimising* $H^R_{1/2}$ *is mathematically identical to maximising* $H^U_{1/2}$*. For any* $\alpha \neq 1/2$*, the weighting of the bounds becomes asymmetric, causing the utility-based and regret-based decisions to diverge, as illustrated in Figure 1.*

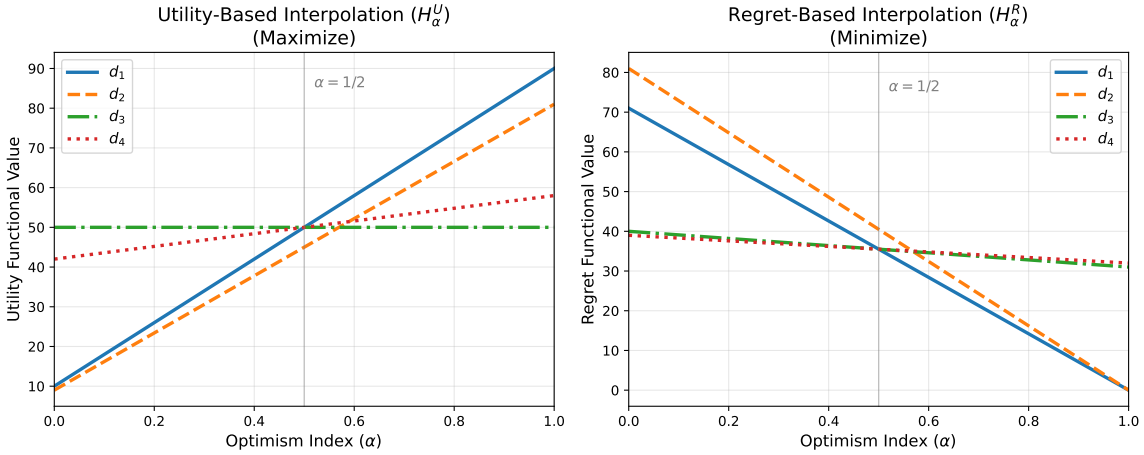

Figure 1: The effect of optimism interpolation in the example. $\alpha = 0$ and $\alpha = 1$ represent the "pure" decision criteria in Table 1.

Choosing α in practice. The optimism index α is not a quantity to be estimated from data; like the utility function itself, it is a subjective input that encodes the decision-maker's attitude towards epistemic ambiguity and should be *elicited* rather than tuned for predictive accuracy. Three reference points anchor the interpolation: $\alpha = 0$ is the maximally

pessimistic Wald stance, which judges each action by its worst-case bound and is appropriate when the downside is catastrophic or irreversible (e.g., patient harm, insolvency, or safety-critical control); $\alpha = 1$ is the maximally optimistic maximax stance, suited to exploratory, opportunity-rich settings in which the cost of a missed opportunity dominates that of an occasional bad outcome; and $\alpha = 1/2$ is the ambiguity-neutral midpoint at which, by Remark 6, the utility and regret criteria coincide. When a practitioner cannot commit to a single value, we recommend reporting decisions across a small grid, such as $\alpha \in \{0, \frac{1}{2}, 1\}$, as a sensitivity analysis: agreement across the grid indicates that the recommendation is robust to the ambiguity attitude, whereas disagreement signals that the outcome hinges on a genuinely subjective judgement that must then be made explicitly. As the empirical results in Section 6 show, the most effective α is problem-dependent and, for a fixed problem, can shift with the sample size as epistemic ambiguity shrinks.

4.2. Definition of the Venn-PDMS

We can now formally define the Venn predictive decision-making system. Mirroring the standard expected utility maximiser from Section 3, a Venn PDMS selects a decision that optimises the chosen ambiguity functional.

Definition 7 (α-utility Venn-PDMS) *Given a multiprobability prediction P and a tolerance $\epsilon > 0$, an* α-utility Venn-PDMS *recommends a decision $F^U_\alpha(z_1, \ldots, z_{n-1}, x_n) \in \boldsymbol{D}$ such that*

$$H^U_\alpha(x_n, F^U_\alpha(z_1, \ldots, z_{n-1}, x_n)) \geq \sup_{d \in \boldsymbol{D}} H^U_\alpha(x_n, d) - \epsilon.$$

Definition 8 (α-regret Venn-PDMS) *Given a multiprobability prediction P and a tolerance $\epsilon > 0$, an* α-regret Venn-PDMS *recommends a decision $F^R_\alpha(z_1, \ldots, z_{n-1}, x_n) \in \boldsymbol{D}$ such that*

$$H^R_\alpha(x_n, F^R_\alpha(z_1, \ldots, z_{n-1}, x_n)) \leq \inf_{d \in \boldsymbol{D}} H^R_\alpha(x_n, d) + \epsilon.$$

Remark 9 *If the decision space $\boldsymbol{D}$ is finite or compact with appropriate continuity conditions on U, the supremum and infimum are attained, allowing us to set $\epsilon = 0$. The PDMSs then simplify to strict optimisers:*

$$F^U_\alpha \in \arg\max_{d \in \boldsymbol{D}} H^U_\alpha(x_n, d) \quad \textit{and} \quad F^R_\alpha \in \arg\min_{d \in \boldsymbol{D}} H^R_\alpha(x_n, d).$$

4.3. Theoretical Guarantees

Any coherent decision-making system operating on von Neumann–Morgenstern utilities must strictly respect the axioms of that framework (see Remark 5 for a discussion about regret-based criteria). Chief among these is the requirement that decisions remain invariant to positive affine transformations, ensuring that the system treats utility as a true interval scale. Furthermore, as a predictive system, it must recover optimal behaviour as epistemic ignorance vanishes. We demonstrate that the Venn–PDMS strictly satisfies these two foundational requirements.

Proposition 10 (Invariance to Affine Transformations) *Fix constants $a > 0$ and $b \in \mathbb{R}$, and let $U'(x, y, d) = aU(x, y, d) + b$ be the corresponding positive affine transformation of the utility function. For any test object x_n and any optimism index $\alpha \in [0, 1]$, the decisions recommended by the α-utility and α-regret Venn-PDMSs under U' are identical to those recommended under U.*

Proof Under U', the state-specific expected utility becomes $E'_i(x_n, d) = aE_i(x_n, d) + b$. Because the maximum operator is translation invariant and scalable by positive constants, the transformed regret is $R'_i(x_n, d) = \max_{d'}(aE_i(d') + b) - (aE_i(d) + b) = aR_i(x_n, d)$. Substituting these into the Hurwicz functionals yields $H^{\mathrm{U}'}_\alpha(x_n, d) = aH^{\mathrm{U}}_\alpha(x_n, d) + b$ and $H^{\mathrm{R}'}_\alpha(x_n, d) = aH^{\mathrm{R}}_\alpha(x_n, d)$. Since $a > 0$, scalar multiplication and constant shifting do not alter the ordering of the decisions. Therefore, the arg max and arg min sets remain the same. ■

In statistical decision theory, the theoretical gold standard is the Bayes-optimal decision, which is the decision that would be made if the true data-generating conditional probability $Q_{Y|X}(\cdot \mid x_n)$ was perfectly known. If we denote this true underlying probability as $p^* := Q_{Y|X}(1 \mid x_n)$, the Bayes-optimal decision simply maximises the standard expected utility: $\arg\max_{d \in \boldsymbol{D}}\{p^*U(x_n, 1, d) + (1 - p^*)U(x_n, 0, d)\}$. The following theorem demonstrates that if the underlying Venn predictor is universal (Vovk et al., 2022, Section 5.3), the decisions made by the Venn-PDMS organically converge to this theoretical optimum as the sample size grows, thereby eliminating the penalty of epistemic ambiguity.

Proposition 11 (Asymptotic Optimality) *Suppose that the object space $\boldsymbol{X}$ is a standard Borel space. Let the Venn-PDMS be constructed using a universal Venn predictor. As the size of the training sequence $n \to \infty$, the decision output by any α-utility or α-regret Venn-PDMS converges in probability to the standard Bayes-optimal decision under the true data-generating distribution, regardless of the choice of decision criterion α.*

Proof By Vovk et al. (2022, Theorem 6.4), for a universal Venn predictor, the maximum variation distance between the multiprobability predictions P_n and the true regular conditional probability $Q_{Y|X}(\cdot \mid x_n)$ converges to 0 in probability as $n \to \infty$. In the binary setting, this variation distance is $\rho(P_n, Q_{Y|X}) = \sum_{y \in \{0,1\}} |P_n\{y\} - Q_{Y|X}(y \mid x_n)|$. For this sum to vanish, we must have both $p^0 \to Q_{Y|X}(1 \mid x_n)$ and $p^1 \to Q_{Y|X}(1 \mid x_n)$. Consequently, the epistemic ambiguity $\Delta p = |p^1 - p^0|$ collapses to 0. When $p^0 = p^1 = p^*$, the upper and lower expected utilities become identical: $\underline{E}(x_n, d) = \overline{E}(x_n, d) = E_{p^*}(x_n, d)$. The ambiguity index α mathematically factors out, reducing the α-utility functional to standard expected utility maximisation. Furthermore, the state-specific regret collapses to the standard expected regret: $R_{p^*}(x_n, d) = \max_{d' \in \boldsymbol{D}} E_{p^*}(x_n, d') - E_{p^*}(x_n, d)$. Because the first term is a constant supremum over the decision space, minimising this regret with respect to d is strictly equivalent to maximising the expected utility $E_{p^*}(x_n, d)$. Thus, both the α-utility and α-regret functionals ultimately yield the identical Bayes-optimal decision. ■

5. The Limits of Probability Aggregation in Predictive Decision Making

A common heuristic when deploying multiprobability predictors (such as Venn–Abers) is to aggregate the multiprobability prediction (p^0, p^1) into a single point estimate p^* and apply standard expected utility (EU) maximisation. A common choice is $p^* = p_{\log} := p^1/(1 - p^0 + p^1)$, which is the minimax regret solution when forecasts are evaluated using the logarithmic loss function (Vovk and Petej, 2012). It is important to realise that this particular aggregation yields optimal decisions if and only if the utility function is logarithmic and the decision criterion is minimax *realised* regret (evaluated ex-post, after the true label is revealed). That is, it assumes the continuous decision space $\boldsymbol{D} = [0, 1]$ and the utility function:

$$U(y, d) = \begin{cases} \log(1 - d) & \text{if } y = 0 \\ \log(d) & \text{if } y = 1. \end{cases}$$

Because our proposed 0-Regret Venn-PDMS evaluates *expected* regret (ex-ante) rather than realised regret, applying $p_{\log}$ even under this specific logarithmic utility function is mathematically sub-optimal. For any other combination of the decision space, utility function, and decision criterion, $p_{\log}$ is similarly invalid. Moreover, for certain combinations of decision space, utility function, and decision criterion, the optimal decision cannot be recovered using *any* aggregated point estimate $p^* \in [p^0, p^1]$, as illustrated in the following example.

Consider a decision space with three actions $\boldsymbol{D} = \{d_1, d_2, d_3\}$, representing two extreme commitments and one conservative hedging action, respectively. We define a simple utility function $U(y, d)$ to be maximised as follows:

- $U(0, d_1) = 10, \quad U(1, d_1) = 0$
- $U(0, d_2) = 0, \quad U(1, d_2) = 10$
- $U(0, d_3) = 4, \quad U(1, d_3) = 4$

The Expected Utility (EU) for any point-estimate probability $p \in [0, 1]$ is:

- $\mathrm{EU}(d_1, p) = 10(1 - p)$
- $\mathrm{EU}(d_2, p) = 10p$
- $\mathrm{EU}(d_3, p) = 4$

Scenario 1: Expected Utility Maximisation under an Aggregated Point Estimate p^* Suppose a practitioner attempts to aggregate a multiprobability prediction into any point estimate $\hat{p} \in [0, 1]$ and applies standard EU maximisation.

- If $\hat{p} < 0.5$, $\mathrm{EU}(d_1, \hat{p}) > 5$. Because $5 > 4$, d_1 strictly dominates d_3.
- If $\hat{p} > 0.5$, $\mathrm{EU}(d_2, \hat{p}) > 5$. Because $5 > 4$, d_2 strictly dominates d_3.
- If $\hat{p} = 0.5$, $\mathrm{EU}(d_1, 0.5) = 5$ and $\mathrm{EU}(d_2, 0.5) = 5$. Both dominate d_3.

Therefore, under standard point-estimate EU maximisation, the hedging action d_3 is strictly dominated across the entire probability space. There exists no aggregated scalar $\hat{p}$ that will trigger d_3.

Scenario 2: Maximin Utility under the Multiprobability Prediction Suppose that the Venn-Abers predictor outputs a multiprobability prediction with high epistemic ambiguity: $(p^0, p^1) = (0.2, 0.8)$. Instead of aggregating, we apply Wald's Maximin Utility criterion (maximising the worst-case utility) directly over the un-aggregated prediction:

- $\min_{p\in\{0.2,0.8\}} \mathrm{EU}(d_1, p) = 10(1 - 0.8) = 2$
- $\min_{p\in\{0.2,0.8\}} \mathrm{EU}(d_2, p) = 10(0.2) = 2$
- $\min_{p\in\{0.2,0.8\}} \mathrm{EU}(d_3, p) = 4$

Under the maximin criterion, the worst-case utility of the hedging action is strictly greater than the worst-case utilities of the others. The optimal decision is unambiguously d_3.

This example illustrates that for certain utility functions, the aggregation and decision operators do not commute. No aggregation function $f(p^0, p^1) \to p^*$ can reproduce the decision behaviour of the interval-based criterion.

Nevertheless, if and when the chosen combination of decision space, utility function, and decision criterion admits an aggregation that yields the same decisions, this can be computationally convenient (e.g., by replacing a non-smooth optimisation problem with a smooth one). A full mathematical characterisation of the topological conditions that permit such aggregation, such as strict concavity in continuous spaces via subdifferential calculus, is left for future work.

6. Example applications

We present two example applications of the various Venn–PDMSs.[2]

Mushroom. Our first example is the mushroom dataset used in (Vovk and Bendtsen, 2018). The mushroom dataset is publicly available from the UCI machine learning repository (mus, 1981). It consists of 8124 observations of whether a mushroom is edible based on 22 attributes. Following Vovk and Bendtsen (2018), we use a utility function that severely punishes eating a poisonous mushroom (shown in Table 2). We employ two Venn-PDMSs,

Decision ($d \in \boldsymbol{D}$)	**Safe** ($y = 0$)	**Poisonous** ($y = 1$)
d_1: Eat	1	-10
d_2: Not eat	0	1

Table 2: Skewed utility function that severely punishes eating a poisonous mushroom.

based on the 1 Nearest Neighbour Venn Machine whose taxonomy is given by

$$\kappa_i = y_j, \quad j := \arg\min_{j\neq i}\{d_H(x_i, x_j)\},$$

where d_H is the Hamming distance between x_i and x_j. Ties in the nearest neighbour are broken by selecting the example with the smallest index. We employ both the 0-utility

2. Code to reproduce all experiments, figures, and tables is available at https://github.com/egonmedhatten/venn-pdms.

Venn-PDMS (Wald) and the 0-regret Venn-PDMS (Savage). These are compared to the 1 nearest neighbour CPDMS described in Vovk and Bendtsen (2018), and the baseline of simply eating only mushrooms predicted to be edible by the 1 nearest neighbour prediction.

To make explicit the difference between the decision criteria, let $\overline{p} = \max\{p^0, p^1\}$ and $\underline{p} = \min\{p^0, p^1\}$. The reader can easily verify that the 0-utility Venn-PDMS recommends eating a mushroom if and only if $1 > 11\overline{p} + \underline{p}$, whereas the 0-regret PDMS recommends eating if and only if $1 > 6(\overline{p} + \underline{p})$. Thus, even if $\underline{p} = 0$, the former requires $\overline{p} < 1/11$, whereas the latter allows us to eat a mushroom if $\overline{p} < 1/6$. As Berger might say, the "Wald picker" is a true "pessimistic statistician" who would rather starve than face even a 10% worst-case chance of being poisoned. The "Savage picker," however, is simply more afraid of looking stupid in hindsight for missing out on a perfectly good meal.

We drew 1000 balanced training sets with sizes $n \in \{4, 6, \ldots, 20\}$ and evaluated the performance of each training set using random balanced test sets of size 10. These deliberately small training sets isolate the small-sample regime in which epistemic ambiguity, and hence the divergence between the decision criteria, is most pronounced; as n grows, the asymptotic optimality of the Venn-PDMS (Section 4) drives all methods towards the same Bayes-optimal behaviour. The mean utility with a 95% confidence interval per training set size is shown in Figure 2.

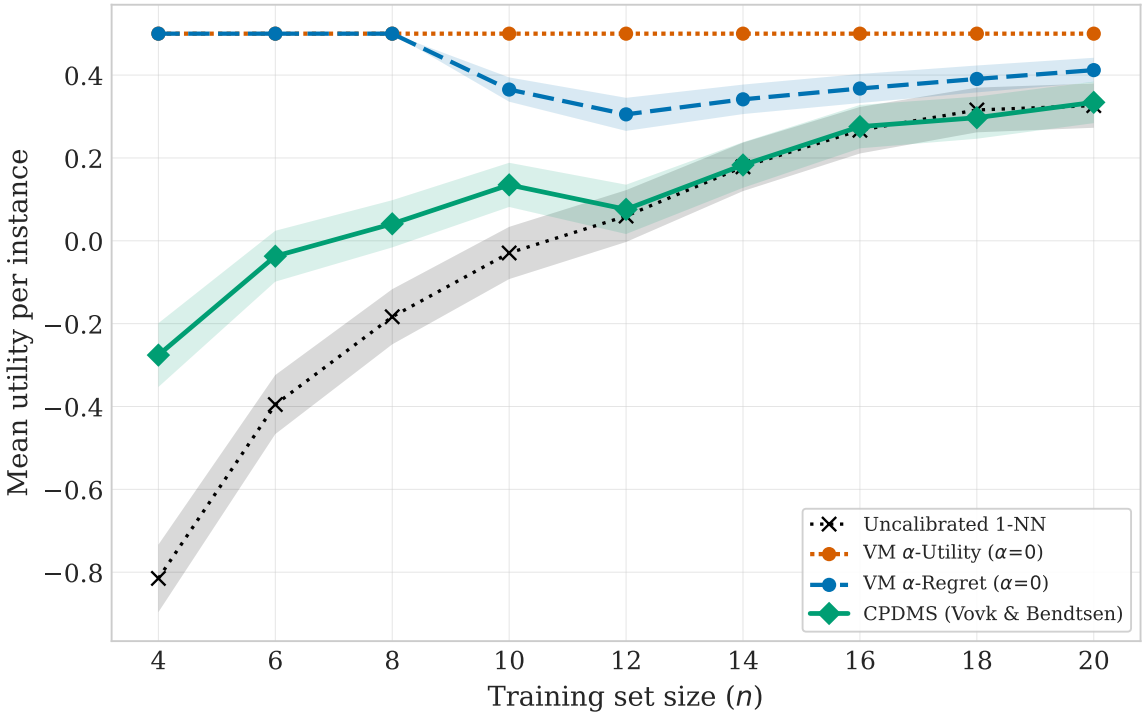

Figure 2: Mean utility with 95% confidence intervals for different training set sizes.

As observed by Vovk and Bendtsen (2018), the CPDMS offsets the decision not to eat potentially poisonous mushrooms but this benefit reduces with increasing training set size. In contrast, the 0-utility Venn-PDMSs exhibit a stable utility of 0.5 for all training set sizes.

Notably, this utility is consistent with a strategy of abstaining from eating any mushrooms at all. The 0-regret Venn-PDMS performs identically to the 0-utility Venn-PDMS for training set sizes smaller than ten, but then its utility drops. Still, it outperforms the CPDMS at all training set sizes.

Figure 3 confirms that the 0-utility Venn-PDMS achieves stable utility by not allowing any mushrooms to be eaten, reflecting its pessimism under epistemic ambiguity. This property keeps us safe under severe epistemic ambiguity and results in a higher utility for all training set sizes than the alternatives. Although standard classification metrics might view this abstention as a trade-off against predictive yield, from a decision-theoretic standpoint, no such trade-off exists. The asymmetric costs are already fully encoded in the utility function. We simply prefer being hungry to being poisoned. An alternative, more complicated, utility function could account for our hunger, explicitly weighing the risk of eating a poisonous mushroom against the risk of starvation; however, this is beyond the scope of this simple example.

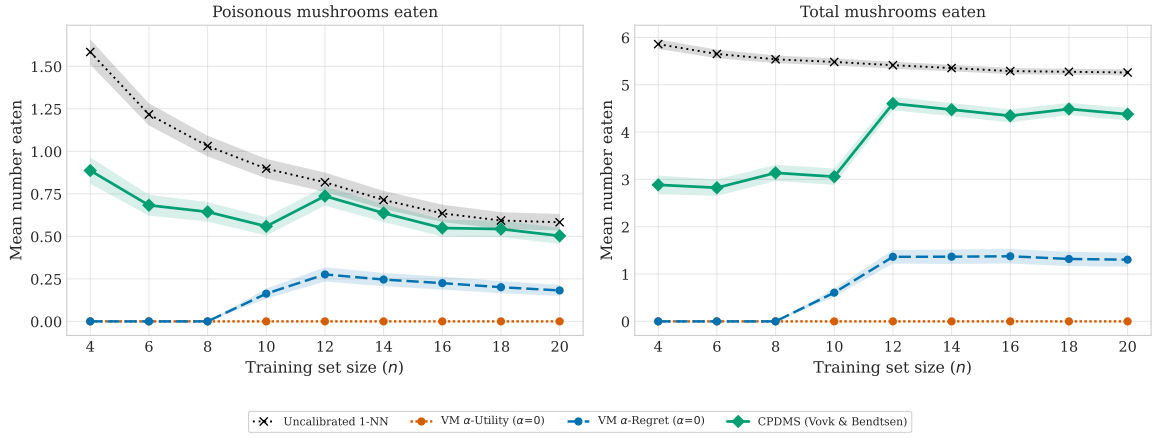

Figure 3: Mean fraction of poisonous mushrooms eaten with 95% confidence intervals (left) and mean number of mushrooms eaten with 95% confidence intervals (right).

German Credit. In the second example, we move from pure classification to a complex real-world financial triage scenario. We evaluated our frameworks using the Statlog German Credit dataset, which is publicly available from the UCI Machine Learning Repository (Hofmann, 1994). The dataset contains 1000 instances describing the creditworthiness of individuals based on 20 attributes, along with the requested loan amount for each instance. To ensure compatibility with distance-based algorithms and gradient boosting, we used the numerical version of the dataset provided by Strathclyde University, which encodes categorical variables as indicators and ordered variables as integers.

Unlike standard classification tasks that treat all errors equally, the original dataset description explicitly mandates the use of an asymmetric cost matrix, noting that approving a bad loan is five times worse than rejecting a good one. We formalise this asymmetry by defining a three-action, amount-dependent utility function, as shown in Table 3. With 20% interest, we follow the dataset description's recommendation.

Decision ($d \in \boldsymbol{D}$)	**Good Credit** ($y = 0$)	**Bad Credit** ($y = 1$)
d_1: Approve	Ar	$-A$
d_2: Review	$(Ar) - 100$	-100
d_3: Reject	0	0

Table 3: Three-action utility function for the German Credit dataset. Payoffs are scaled by the requested loan amount A and the interest rate r. The 'Review' action simulates routing the application to a human underwriter at a fixed cost of 100 DM.

This three-action framework creates a dynamic triage problem. The "Approve" action carries the highest potential profit (a 20% interest return) but exposes the system to catastrophic loss if the borrower defaults. To operationalise the "Review" action (d_2), we assign a fixed manual underwriting cost of 100 DM. This value was chosen as it represents approximately 3% of the mean loan amount (3271 DM) in the 1994 dataset, serving as a highly realistic proxy for the administrative overhead and labour costs of a human-in-the-loop manual review at the time. Crucially, from a decision-theoretic perspective, this fixed cost creates the necessary structural threshold: it is low enough to act as a rational safety net under severe epistemic ambiguity, yet high enough to incentivise automated decisions once the multiprobability bounds sufficiently converge. Therefore, a successful decision framework must learn to dynamically adjust its automation rate: safely routing uncertain cases to human underwriters during the initial "cold start" phase and organically scaling up automation as epistemic ambiguity decreases.

To evaluate the impact of epistemic ambiguity on dynamic triage, we compared standard machine learning approaches with our proposed multiprobability frameworks. For all methods, we extracted a base scoring function (e.g., probability estimates or distance quotients) and evaluated performance across training set sizes $n \in \{10, 25, 50, 100, 200, 250, 500\}$, simulating a cold-start deployment that gradually accumulates data. To obtain stable estimates and quantify run-to-run variability, we repeated the entire protocol over 1000 independent trials. Each trial, indexed by a distinct random seed, draws a fresh stratified partition of the 1000 instances: a held-out test set of 200 applications is set aside, and the training set of size n is sampled, again with stratification, from the remaining 800 instances. All methods within a trial are evaluated on the same test set, and every reported quantity is aggregated over the 1000 trials. We define three baselines that, in different ways, forgo an explicit account of epistemic ambiguity:

1. **Hard Classifier:** This represents the standard operational paradigm where a model outputs a rigid class label, equivalent to a degenerate multiprobability prediction where $p^0 = p^1 \in \{0, 1\}$. By asserting absolute certainty, this baseline mathematically eliminates the possibility of the "Review" action (d_2), forcing the system into a binary choice between "Approve" and "Reject."

2. **Uncalibrated Baseline:** This approach utilizes the continuous raw point estimate $\hat{p}$ from the underlying model to maximize expected utility. While capable of selecting the "Review" action, it treats its point estimate as perfectly precise, thus ignoring the epistemic ambiguity inherent to small training samples.

3. **Inductive CPDMS:** The conformal predictive decision-making system of Vovk and Bendtsen (2018), which we adapt to the inductive (split-calibration) setting and to the three-action utility. Using a held-out calibration set, it forms a conformal predictive distribution of the utility of each action and selects the single action with the highest calibrated expected utility. Because a conformal predictive system collapses its prediction into a single distribution function, it accounts for aleatoric risk but retains no representation of epistemic ambiguity: its "Review" decisions are triggered only when the calibrated expected utility of manual review exceeds that of approval or rejection, and never by a notion of what the model does not know.

Against these baselines, we evaluated α-regret- and α-utility Venn-PDMSs using both the *Venn-Abers predictor* (VAP) and *inductive Venn-Abers predictor* (IVAP) (Vovk and Petej, 2012). We considered $\alpha \in \{0.0, 0.5, 1.0\}$ reflecting different levels of optimism. As the underlying classifier, which was also used as the scoring function for the Venn-Abers predictors, we used the Gradient Boosting Classifier from scikit-learn (Pedregosa et al., 2011; Friedman, 2001) because of its robust, state-of-the-art performance on heterogeneous tabular data typical of financial risk assessment. The training set was randomly split into 67/33 proper training and calibration sets for the IVAP-based Venn PDMSs.

The mean utility gain over a strategy that defers all decisions to a human reviewer is reported in Table 4. Notably, the 0-regret Venn-PDMS using VAP allows for a safe, even profitable ($\sim$128 DM) deployment, even for $n \in \{10, 25, 50, 100\}$. As shown in Table 6, the automation rate at these small training set sizes is small, approximately 2%, because the epistemic ambiguity is large. Thus, the safety is attributable to the Venn-PDMS "knowing what it does not know", safely deferring decisions it cannot handle. The 0-utility Venn-PDMS, in contrast, is completely paralysed by epistemic ambiguity. It rejects all applications to protect against the absolute worst-case outcome, resulting in 100% automation but no business.

The hard classifier performs poorly because it is oblivious to the utility function, whereas the baseline, using a point estimate $\hat{p}$ takes the utility into account but has no notion of epistemic ambiguity. Consequently, it automates aggressively, reaching profitable automation only for $n = 500$. It is interesting to note that the performance of the α-regret Venn-PDMSs decreases with α. Apparently, Berger's pessimistic statisticians (Berger, 1993) were on to something. For the α-utility Venn-PDMSs, $\alpha = 0.5$ has the best overall performance. Because Wald's maximin criterion is maximally pessimistic, protecting against the absolute worst outcome, softening this extreme behaviour may be desirable in situations where catastrophic failure is not a major concern. The IVAP-based Venn-PDMSs initially suffer from data splitting. Their performance increases with larger training set sizes. For $n = 500$, the $\frac{1}{2}$-utility Venn-PDMS using IVAP performs best overall. As Remark 6 states, the utility and regret criteria make the same decisions for $\alpha = 1/2$.

The behaviours are illuminated by the diagnostics in Table 5. The Gradient Boosting scoring function improves steadily with n, its accuracy rising from 0.62 to 0.74 and its Brier

Strategy	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 250$	$n = 500$
Uncalibrated Baseline	-130112.7	-114492.8	-54242.1	-19885.7	-6149.5	-3638.6	+327.1
Hard Classifier	-153622.7	-143994.0	-137488.3	-128918.8	-120010.6	-118225.7	-108779.7
α-Utility Venn-PDMS							
IVAP $\alpha = 0.0$	-10872.0	-6677.3	-6104.2	-3598.0	-2056.0	-1587.7	-546.1
IVAP $\alpha = 0.5$	-215.7	-451.7	-416.8	-178.3	+26.8	+149.8	**+529.8**
IVAP $\alpha = 1.0$	-67032.3	-20337.6	-25675.6	-14296.3	-7026.7	-4677.7	-1706.1
VAP $\alpha = 0.0$	-63908.8	-63908.8	-63908.8	-63908.8	-62711.6	-59318.9	-37093.2
VAP $\alpha = 0.5$	+90.2	+90.2	+90.2	+90.3	+103.2	+73.1	-629.4
VAP $\alpha = 1.0$	-217284.3	-217284.3	-217284.3	-217283.1	-214991.4	-203259.1	-77687.9
α-Regret Venn-PDMS							
IVAP $\alpha = 0.0$	-231.6	-455.8	-419.4	-180.7	+25.9	**+150.1**	+529.1
IVAP $\alpha = 0.5$	-215.7	-451.7	-416.8	-178.3	+26.7	+149.8	+529.7
IVAP $\alpha = 1.0$	-67032.3	-20337.6	-25675.6	-14296.3	-7026.7	-4677.7	-1706.1
VAP $\alpha = 0.0$	**+128.1**	**+128.1**	**+128.1**	**+128.2**	**+142.1**	+128.4	-465.4
VAP $\alpha = 0.5$	+90.2	+90.2	+90.2	+90.3	+103.2	+73.1	-629.4
VAP $\alpha = 1.0$	-217284.3	-217284.3	-217284.3	-217283.1	-214991.4	-203259.1	-77687.9
Inductive CPDMS							
Inductive CPDMS	-67553.6	-56929.9	-47598.3	-24462.3	-10962.3	-8627.0	-3410.9

Table 4: Mean utility gain over Human Reviewer (DM) across training-set sizes (German Credit dataset). Positive values indicate that the strategy is profitable relative to reviewing all applications. Best per column in bold.

score falling from 0.36 to 0.17, but the two calibration schemes react very differently in the cold-start regime. The transductive VAP produces maximally ambiguous predictions ($\Delta p \approx 1$) for $n \leq 100$. This saturation is a genuine property of the full Venn–Abers construction rather than an artefact: for each test object, the scoring function is retrained on the training sample augmented with that object under each tentative label, and at small n this expressive scorer can accommodate the augmented point under either label almost perfectly, so the two labellings are near-equally consistent with the data. The reported width therefore reflects the flexibility of the scoring function as much as the sample size; validity holds regardless of the scorer, and only its efficiency is affected. This maximal ambiguity is precisely why the 0-regret VAP defers almost every application and thereby stays safe. The split-calibrated IVAP instead fixes its scorer, which never sees the tentative label, and so yields much narrower intervals (Δp falling from 0.37 to 0.06) that track the information in the calibration set more directly, at the cost of a noisier base estimate, so its ambiguity-neutral variants can automate earlier. As n grows and the VAP width finally contracts, this strong caution is no longer rewarded: by $n = 500$ the $\alpha = 1/2$ IVAP overtakes the 0-regret VAP, whose gain turns negative (-465 DM). This is exactly the sample-size dependence of the most effective α anticipated in Section 4.1.

Finally, the inductive CPDMS illustrates that the conformal calibration of the point predictor is, by itself, insufficient. Its conformal predictive distribution does temper the overconfidence of the uncalibrated baseline in the extreme cold-start regime, roughly halving the loss at $n = 10$ (-67554 against -130113 DM). However, because it commits to a single

	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 250$	$n = 500$
Base accuracy (GB)	0.619	0.636	0.674	0.699	0.719	0.725	0.745
Brier score (GB)	0.360	0.341	0.270	0.227	0.199	0.192	0.174
Mean width Δp (IVAP)	0.365	0.224	0.211	0.157	0.108	0.094	0.064
Mean width Δp (VAP)	1.000	1.000	1.000	1.000	0.988	0.950	0.666

Table 5: Diagnostics for the German Credit experiment as a function of training-set size n, averaged over the 1000 trials. The base accuracy and Brier score characterise the underlying Gradient Boosting scoring function, while the mean multiprobability width $\Delta p = |p^1 - p^0|$ quantifies the epistemic ambiguity of the IVAP and VAP predictions.

predictive distribution, it automates aggressively; between 27% and 58% of decisions (Table 6), and lacking any epistemic trigger for deferral, it never becomes profitable. It remains at -3411 DM even at $n = 500$, where the uncalibrated baseline has already turned a small profit, and at no training-set size does it approach the safe, profitable operation of the 0-regret VAP Venn-PDMS.

Strategy	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 250$	$n = 500$
Uncalibrated Baseline	90.8	89.1	67.1	48.7	35.6	32.0	24.9
Hard Classifier	100.0	100.0	100.0	100.0	100.0	100.0	100.0
α-Utility Venn-PDMS							
IVAP $\alpha = 0.0$	24.8	19.4	17.9	15.2	13.5	13.1	13.5
IVAP $\alpha = 0.5$	8.3	9.9	10.5	11.4	13.1	13.1	15.2
IVAP $\alpha = 1.0$	38.1	17.6	27.3	24.3	21.7	20.2	20.1
VAP $\alpha = 0.0$	100.0	100.0	100.0	100.0	98.4	94.4	70.8
VAP $\alpha = 0.5$	11.5	11.5	11.5	11.5	11.8	12.9	23.0
VAP $\alpha = 1.0$	100.0	100.0	100.0	100.0	99.4	96.9	73.1
α-Regret Venn-PDMS							
IVAP $\alpha = 0.0$	7.8	9.8	10.4	11.4	13.1	13.1	15.2
IVAP $\alpha = 0.5$	8.3	9.9	10.5	11.4	13.1	13.1	15.2
IVAP $\alpha = 1.0$	38.1	17.6	27.3	24.3	21.7	20.2	20.1
VAP $\alpha = 0.0$	1.8	1.8	1.8	1.8	2.1	3.6	17.5
VAP $\alpha = 0.5$	11.5	11.5	11.5	11.5	11.8	12.9	23.0
VAP $\alpha = 1.0$	100.0	100.0	100.0	100.0	99.4	96.9	73.1
Inductive CPDMS							
Inductive CPDMS	57.8	57.3	53.9	43.1	33.3	31.5	27.0

Table 6: Mean automation rate (%) across training-set sizes (German Credit dataset).

The median utility per training-set size, with a shaded band spanning the 2.5–97.5 percentile range across the 1000 trials, is shown in Figure 4 for a selection of methods. We

plot the median here, rather than the mean reported in the tables, because it is more robust to the heavy-tailed distribution of outcomes in the cold-start regime.

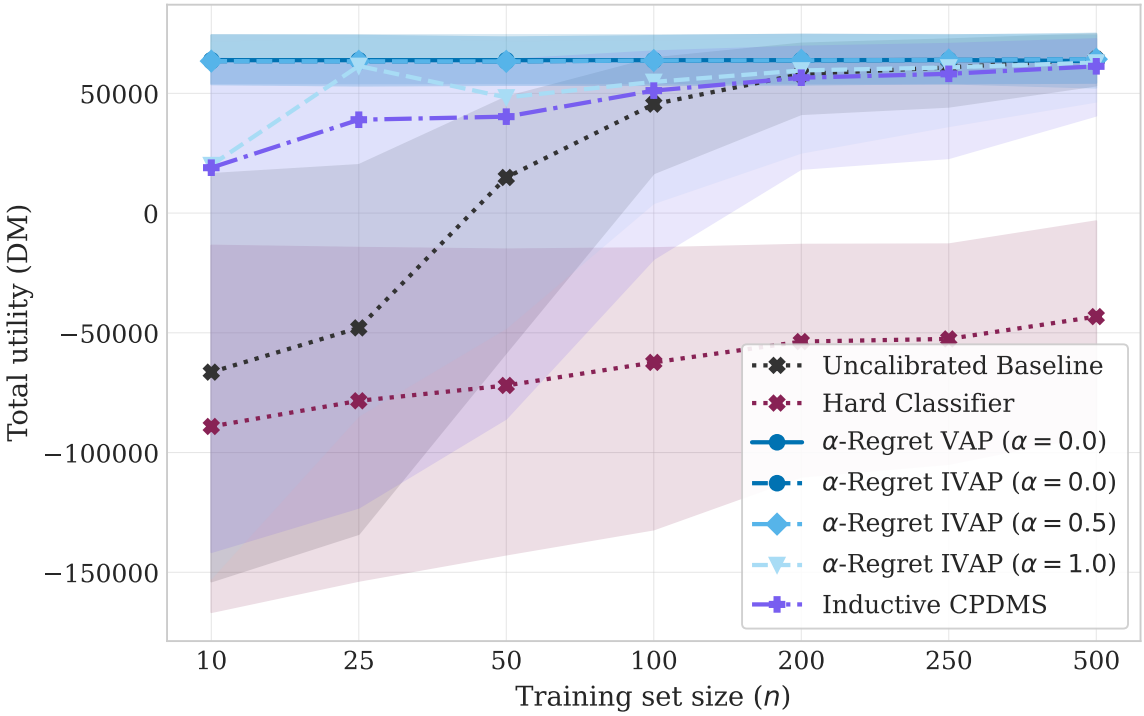

Figure 4: Median utility across 1000 random trials with a shaded band spanning the 2.5–97.5 percentile range, for a selection of methods, including the inductive CPDMS baseline, on the German Credit dataset.

On exchangeability. The calibration guarantees underlying the Venn-Abers and conformal predictors assume exchangeable data. Our protocol satisfies this assumption by construction: each train/calibration/test partition is obtained by uniformly shuffling the dataset, so the resulting samples are exchangeable by design. This is appropriate for isolating the effect of epistemic ambiguity under controlled conditions, but it is an idealisation. In a genuine deployment, credit applications arrive sequentially, and the data-generating distribution is subject to drift—through changing macroeconomic conditions, evolving applicant populations, and revised lending policies—so exchangeability is unlikely to hold exactly. Adapting the Venn-PDMS to non-exchangeable, drifting data is an important direction for practical application in situations where the exchangeability assumption breaks down.

7. Concluding Discussion

We extended the method of Venn prediction to make it applicable to decision-making in a way that explicitly accounts for the epistemic ambiguity (Knightian uncertainty) of the multiprobability predictions. We argued that epistemic ambiguity necessitates a subjec-

tive stance towards decision criteria: are we optimists or pessimists, and do we care about opportunity cost (regret) or absolute utility? Different decision-making scenarios suggest different criteria. In some settings, avoiding catastrophic failure requires a pessimistic consideration of absolute utility. In others, opportunity cost (not wanting to look stupid in hindsight suggests a pessimistic regret criterion) is more important. Yet in exploratory, opportunity-rich settings—for instance, cheaply screening a large pool of candidates where a missed opportunity is more costly than an occasional bad draw—an optimistic stance is instead warranted. Our framework allows users to interpolate between strict pessimism and strict optimism using an optimism index α.

We compared two of our Venn-PDMSs with the Conformal Predictive Decision-Making System (CPDMS) (Vovk and Bendtsen, 2018) using the mushroom dataset. Our methods consistently outperformed the CPDMS; however, this is a limited example on a simple dataset. Further empirical and theoretical comparisons are necessary to ascertain the conditions under which Venn-PDMSs may surpass CPDMSs, and vice versa. Nonetheless, one can observe certain theoretical and computational advantages in our proposed framework. The Venn-PDMS requires only a single multiprobability prediction per test object x, whereas the original CPDMS formulation (Vovk and Bendtsen, 2018) computes one conformal predictive distribution per decision, utilising a training sequence transformed into the utility space $(x_1, U(y_1, d)), x_2, U(y_2, d), \dots)$. As noted by Lanngren et al. (2025), it is feasible to circumvent this computational bottleneck by employing the same CPD, derived from the original training sequence, to compute expected utility. The primary advantage of the Venn-PDMS, however, is theoretical. In scenarios where epistemic uncertainty is significant, our framework explicitly addresses it through subjective decision criteria under uncertainty. This adaptability enables decision-support systems to "know what they do not know," allowing them to safely refrain from automation when uncertainty is excessive. This was exemplified in the German Credit case, where some Venn-PDMSs achieved profitable, low-level automation even with severely limited data.

Interestingly, epistemic uncertainty is present in conformal predictive systems; the CPD is obtained by random interpolation between a lower and upper predictive distribution: a CPD $\Pi(y, \tau)$ can be represented as $(1-\tau)\Pi(y, 0) + \tau\Pi(y, 1)$. Future work could investigate whether this epistemic ambiguity (informally, $\Pi(y, 1) - \Pi(y, 0)$) can be used for uncertainty-aware CPDMs.

We hope that Venn–PDMSs will prove useful in practice. Empirical investigations of the framework appear to be the most important direction for future research, as our examples here are limited in scope. In particular, our finding that a more pessimistic stance is advantageous is specific to these datasets and utility structures and should not be read as a general prescription; larger and more challenging benchmarks are needed to map out when each criterion excels. We plan to extend the evaluation to higher-dimensional, structured problems such as molecular mutagenicity prediction (the Ames test) and to a broader set of baselines, including Bayesian decision-theoretic methods. Although we developed the framework for binary labels, the multiprobability construction and the α-criteria extend to larger finite label spaces; a systematic treatment of the multiclass setting is a natural next step. Other directions include, as mentioned in Section 5, characterising when aggregating multiprobability predictions leads to equivalent decision-making, particularly in continuous decision spaces such as drug doses, as this may be significantly more computationally

convenient. Continuing the theme of ambiguity, it may sometimes be difficult to exactly describe a utility function, but rather easy to provide upper and lower utility bounds given an outcome. Extending the framework for *imprecise utilities*, as in Seidenfeld et al. (1995); Jansen et al. (2018), may be an interesting future direction. Finally, the decision criteria presented in this note are not the only possible ones. The field of imprecise probability (IP) has several decision criteria, including E-admissibility and maximality (Augustin et al., 2014, Chapter 8), that may complement our current framework.

References

Mushroom. UCI Machine Learning Repository, 1981. DOI: https://doi.org/10.24432/C5959T.

Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1993.

Herman Chernoff. Rational selection of decision functions. *Econometrica*, 22(4):422–443, 1954. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907435.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/1884324.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2699986.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989. ISSN 0304-4068. doi: https://doi.org/10.1016/0304-4068(89)90018-9. URL https://www.sciencedirect.com/science/article/pii/0304406889900189.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Leonid Hurwicz. Optimality criteria for decision making under ignorance. Technical report, Cowles Commission Discussion Paper, Statistics, No. 370, 1951.

C. Jansen, G. Schollmeyer, and T. Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131, 2018. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2018.04.011. URL https://www.sciencedirect.com/science/article/pii/S0888613X1730573X.

Frank Hyneman Knight. *Risk, uncertainty and profit*, volume 31. Houghton Mifflin, 1921.

Simon Lanngren, Martin Toremark, and Johan Hallberg Szabadváry. Conformal predictive decision making: A comparative study with bayesian and point-predictive methods. In Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson, and Henrik Boström, editors, *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pages 379–404. PMLR, 10–12 Sep 2025. URL https://proceedings.mlr.press/v266/lanngren25a.html.

R. Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. Wiley, New York, 1957.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598783.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Leonard Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Leonard J Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951.

Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. A representation of partially ordered preferences. *The Annals of Statistics*, 23(6):2168 – 2217, 1995. doi: 10.1214/aos/1034713653. URL https://doi.org/10.1214/aos/1034713653.

J Von Neumann and O Morgenstern. *Theory of games and economic behavior, 2nd rev.* Princeton University Press, 1947.

Vladimir Vovk. Well-calibrated predictions from online compression models. In Ricard Gavaldá, Klaus P. Jantke, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 268–282, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-39624-6.

Vladimir Vovk and Claus Bendtsen. Conformal predictive decision making. In *Conformal and Probabilistic Prediction and Applications*, pages 52–62. PMLR, 2018.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World, Second Edition*. Springer International Publishing, 1 2022. ISBN 9783031066498. doi: 10.1007/978-3-031-06649-8/COVER.

Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950.