Optimal Conformal Prediction under Epistemic Uncertainty

Alireza Javanmardi[1,2] **Soroush H. Zargarbashi**[3] **Santo M. A. R. Thies**[1,2]
Willem Waegeman[4] **Aleksandar Bojchevski**[5] **Eyke Hüllermeier**[1,2,6]

[1]LMU Munich [2]MCML [3]CISPA [4]Ghent University [5]University of Cologne [6]DFKI

Abstract

Conformal prediction (CP) is a widely used frequentist framework to quantify uncertainty by constructing prediction sets with user-specified marginal coverage guarantees. In practice, CP is typically applied on top of probabilistic classifiers, which are able to express aleatoric but not epistemic uncertainty. In this paper, we consider the question of how to optimally employ CP on top of a more expressive formalism, namely credal sets, which can express both aleatoric and epistemic uncertainty. More specifically, we propose probabilistic Bernoulli prediction sets (BPS) and derive a variant that achieves conditional coverage for valid credal sets while remaining minimal in expected size. We then address the more realistic scenario in which the validity of the credal sets is not guaranteed. Assuming access to calibration data with ground-truth distributions over labels, we apply conformal risk control to BPS and derive a PAC-style guarantee: with high probability over the data, the achieved conditional coverage is at least the desired level. We validate our theoretical findings empirically over various datasets.

1 INTRODUCTION

Modern neural networks are widely deployed in safety-critical applications, where reliable uncertainty quantification is essential. In practice, uncertainty is typically modeled probabilistically, for instance, using a probabilistic classifier (e.g., a neural network with a softmax output), which predicts a label distribution for each input. Not only are such predictions often miscalibrated [Guo et al., 2017], but they also fail to distinguish between two distinct sources of uncertainty, commonly known as *aleatoric* and *epistemic* uncertainty [Hora, 1996, Hüllermeier and Waegeman, 2021]. Aleatoric uncertainty arises from inherent randomness in the data and is irreducible, whereas epistemic uncertainty reflects the learner's lack of knowledge and can, in principle, be reduced. More expressive representations of uncertainty that are able to capture both aleatoric and epistemic uncertainty exist in two main forms: credal sets, which are convex sets of plausible probability distributions over labels [Walley, 1991, Zaffalon, 2002, Caprio et al., 2024b,a, Wang et al., 2024, Löhr et al., 2025, Nguyen et al., 2025, Hofman et al., 2026], and second-order distributions, which are distributions over label distributions [Neal, 2012, Blundell et al., 2015, Daxberger et al., 2021, Sensoy et al., 2018]. Despite their appealing properties, such representations lack a formal notion of reliability in the uncertainties they express.

Formal reliability guarantees are provided by conformal prediction (CP) [Vovk et al., 2022], a framework for uncertainty representation based on set-valued predictions: rather than returning a single label, CP outputs a set of plausible labels. Such prediction sets are practical objects for risk-averse decision-making, as highlighted by Kiyani et al. [2025]. Importantly, CP guarantees marginal coverage, that is, on average, over a random draw of the data, the true label is contained in the predicted set with the user-specified probability. This guarantee is distribution-free, holds in finite samples, and is agnostic to the underlying predictive model, requiring only a set of holdout calibration data that are exchangeable with the future test data. Conformal risk control (CRC) [Angelopoulos et al., 2024a] proposes similar distribution-free guarantees to control an arbitrary user-defined risk function — it ensures that risk is upper bounded for the future test point. CP is a particular case of CRC with set-valued prediction and miscoverage risk.

In practice, CP is most commonly applied on top of probabilistic classifiers and has demonstrated strong empirical performance across a wide range of classification tasks [Sadinle et al., 2019, Romano et al., 2020, Angelopoulos et al., 2021]. The core component of CP is the nonconformity score, a function that quantifies the (im)plausibility of assigning a candidate label to a given input. When com-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2461–2479.

bined with probabilistic classifiers, this score is typically defined as a function of the predicted label distribution, e.g., the negative predicted probability (softmax) of the label [Sadinle et al., 2019]. Among various approaches, Romano et al. [2020] introduced the Adaptive Prediction Sets (APS) method, which, given access to the oracle label distributions, returns the smallest possible prediction sets that achieve conditional coverage, i.e., that contain the true label with a user-specified probability for each input.

In this paper, we are interested in constructing reliable prediction sets à la CP when uncertainty is represented by a credal set predictor rather than by a probabilistic classifier. One can look at this problem from two different perspectives. From the point of view of epistemic uncertainty representation, the question is how to "conformalize" a credal set predictor so as to deliver reliable set predictions. From the point of view of conformal prediction, the question is how to benefit from the additional information about epistemic uncertainty provided by a credal set predictor compared to a probabilistic classifier (see Appendix A for a discussion of related approaches at this intersection).

As one important benefit, let us highlight guaranteed conditional coverage. Imagine, for example, a credal set predictor is able to guarantee that the conditional probability of label A given query instance $\boldsymbol{x}$ is between 0.6 and 0.7, the probability of label B is between 0.3 and 0.4, and the probability of C is between 0 and 0.1. The learner can then be sure that the prediction set $\{A, B\}$ has a conditional coverage of (at least) 0.9, i.e., that $\{A, B\}$ is a valid prediction if the (user-specified) error rate is 0.1. This guarantee could not be given on the basis of any first-order representation, e.g., a representative label distribution of the credal set.

We approach the problem step by step, progressing from simpler to more difficult variants. We begin with the setting of valid credal sets, where the true label distribution is guaranteed to lie within the credal set for each input (like in the previous example). In this setting, we introduce Bernoulli Prediction Sets (BPS), which take a credal set as input and return the smallest prediction set satisfying conditional coverage at any given desired level. As the assumption of valid credal predictions is unrealistic in practice, we first consider the problem of optimal set prediction under a suitably relaxed assumption of partially valid credal sets, and finally, in a scenario where no validity properties can be guaranteed at all. In these latter two cases, we obtain a probably approximately correct (PAC)-style conditional coverage guarantee: with high probability, the achieved conditional coverage is at least the desired level.

We evaluate our proposed methods on multiple datasets under settings with valid, partially valid, and unknown-validity credal sets, demonstrating the practical relevance of our theoretical results.

2 BACKGROUND

We assume that inputs $\boldsymbol{x}_i$ are sampled i.i.d. from a distribution $\mathcal{P}_{\mathcal{X}}$. Conditional on $\boldsymbol{x}_i$, the label y_i is drawn from the categorical distribution $p(\cdot \mid \boldsymbol{x}_i) \in \triangle^K$, where K is the number of classes and $\triangle^K$ is the $(K-1)$-dimensional probability simplex. In this setting, we write $\boldsymbol{p}_i := p(\cdot \mid \boldsymbol{x}_i)$ and refer to it as the oracle label distribution at $\boldsymbol{x}_i$ (which is uniquely determined by $\boldsymbol{x}_i$). Even if the true distribution $\boldsymbol{p}_i$ were known, the prediction of the label y_i is subject to aleatoric uncertainty due to the inherent randomness in sampling $y_i \sim \boldsymbol{p}_i$. For a given input $\boldsymbol{x}_i$, a probabilistic classifier outputs a single probability distribution over labels, denoted by $\boldsymbol{\pi}_i = \pi(\cdot \mid \boldsymbol{x}_i) \in \triangle^K$, which serves as an estimate of $\boldsymbol{p}_i$. While such an estimate can represent aleatoric uncertainty, it cannot capture the discrepancy between $\boldsymbol{\pi}_i$ and $\boldsymbol{p}_i$, which corresponds to epistemic uncertainty. A credal set predictor captures this discrepancy by outputting a convex set of plausible probability distributions over labels, denoted by $\mathcal{Q}_i := \mathcal{Q}(\boldsymbol{x}_i) \subseteq \triangle^K$. Roughly speaking, the larger the credal set, the greater the learner's uncertainty about the true distribution $\boldsymbol{p}_i$.

In the standard classification setting, one typically observes *zero-order* data consisting of pairs $(\boldsymbol{x}_i, y_i)$, where each input $\boldsymbol{x}_i$ is paired with a single realized label y_i. In a richer setting, one may instead have access to *first-order* data consisting of pairs $(\boldsymbol{x}_i, \boldsymbol{p}_i)$, where $\boldsymbol{x}_i$ is paired with the true label distribution $\boldsymbol{p}_i$. Such first-order data can arise in practice, for example, by aggregating multiple annotations per instance into a label distribution [Peterson et al., 2019, Nie et al., 2020, Obuchowicz et al., 2020, Schmarje et al., 2022], and are therefore becoming relevant in a growing range of applications, including conformal prediction [Stutz et al., 2023, Javanmardi et al., 2024, Caprio et al., 2025].

Conformal prediction and risk control. Let $\mathcal{D}_{\text{cal}}^{\text{zero}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a holdout zero-order calibration set exchangeable with the future test point $(\boldsymbol{x}_{n+1}, y_{n+1})$. For any nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (capturing the disagreement between $\boldsymbol{x}$ and y), and any user-specified coverage rate $1-\alpha$, conformal prediction [Vovk et al., 2022] constructs sets $\mathcal{C}(\boldsymbol{x}_{n+1})$ as follows. Let $\mathbb{Q}\left(1-\alpha; \mathcal{A}\right)$ denote the $(1-\alpha)\cdot(1+\frac{1}{n})$ quantile of the set $\mathcal{A}$ and let $q := \mathbb{Q}\left(1-\alpha; \{s(\boldsymbol{x}_i, y_i) : (\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{cal}}^{\text{zero}}\}\right)$. Then,

$$\mathcal{C}(\boldsymbol{x}_{n+1}) = \{y : s(\boldsymbol{x}_{n+1}, y) \leq q\}\,. \tag{1}$$

In the common classification setting, the nonconformity score is often defined in terms of the predicted label distribution of a probabilistic classifier; for example, $s(\boldsymbol{x}, y) = -\pi(y \mid \boldsymbol{x})$ [Sadinle et al., 2019]. CP guarantees

$$\mathrm{P}_{\mathcal{D}_+}[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{n+1})] \geq 1-\alpha, \tag{2}$$

with $\mathcal{D}_+ := \mathcal{D}_{\text{cal}}^{\text{zero}} \cup \{(\boldsymbol{x}_{n+1}, y_{n+1})\}$. Moreover, when the scores have a continuous joint distribution, and there are no

ties, the coverage probability is upper bounded by $1-\alpha+\frac{1}{n+1}$ [Lei et al., 2018]. This guarantee holds only marginally, i.e., on average over a random draw of calibration and test data. In contrast, conditional coverage requires the same coverage probability at every test point, i.e.,

$$\mathrm{P}\left[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{n+1}) \mid \boldsymbol{x}_{n+1}\right] \geq 1-\alpha. \tag{3}$$

Generalizing CP, one can provide a similar guarantee for arbitrary risk functions through conformal risk control (CRC) [Angelopoulos et al., 2024a]. Consider a family of set predictors $\mathcal{C}_\lambda$ indexed by a parameter λ, and let $\mathcal{L}_i(\lambda) := \mathcal{L}(\mathcal{C}_\lambda(\boldsymbol{x}_i), y_i)$ denote the risk induced by the set $\mathcal{C}_\lambda(\boldsymbol{x}_i)$. For exchangeable $\mathcal{D}_+$, if $\mathcal{L}(\lambda)$ is non-increasing in λ, bounded above (i.e., $\mathcal{L}(\lambda) \in (-\infty, b]$), and right continuous in λ, Angelopoulos et al. [2024a] show:

$$\mathbb{E}_{\mathcal{D}_+}[\mathcal{L}_{n+1}(\lambda^\star)] \leq \alpha \tag{4}$$

for

$$\lambda^\star := \inf\left\{\lambda : \frac{1}{n+1}\sum_{i=1}^{n}\mathcal{L}_i(\lambda) + \frac{b}{n+1} \leq \alpha\right\}.$$

Intuitively, λ is a conservativeness parameter: increasing λ typically produces larger (more conservative) prediction sets and therefore reduces the risk. CRC selects the smallest λ that ensures the expected risk remains below the predefined tolerance level α. Note that conformal prediction itself is a special case of CRC with $\mathcal{C}_\lambda(\boldsymbol{x}_i) = \{y : s(\boldsymbol{x}_i, y) \leq \lambda\}$, where the risk is defined as the miscoverage; that is, $\mathcal{L}_i(\lambda) = \mathbb{1}[y_i \notin \mathcal{C}_\lambda(\boldsymbol{x}_i)] = \mathbb{1}[s(\boldsymbol{x}_i, y_i) > \lambda]$.

Adaptive prediction sets. CP provides a marginal guarantee, meaning it is possible to have lower coverage in some regions of $\mathcal{X}$ and higher coverage in others. For instance, with $s(\boldsymbol{x}, y) = -\pi(y \mid \boldsymbol{x})$, the coverage probability can be biased toward easy examples (i.e., cases with highly concentrated predicted distributions). To address this, Adaptive Prediction Sets (APS) [Romano et al., 2020] use a score function that aims to better approximate conditional coverage. Formally, the APS score is defined as $s(\boldsymbol{x}, y) := \rho(\boldsymbol{x}, y) + u \cdot \pi(y \mid \boldsymbol{x})$, where $\rho(\boldsymbol{x}, y) := \sum_{y' \in \mathcal{Y}} \pi(y' \mid \boldsymbol{x}) \mathbb{1}[\pi(y' \mid \boldsymbol{x}) > \pi(y \mid \boldsymbol{x})]$ is the cumulative probability of classes ranked above y, and $u \sim \mathrm{Uniform}[0,1]$ is a tie-breaking random variable used to achieve exact $1-\alpha$ coverage.[1] Given access to the oracle label distributions, APS produces the smallest prediction sets satisfying conditional coverage [Angelopoulos et al., 2024b]. Notably, even without such oracle access, APS tends to distribute coverage more evenly across test inputs than other CP baselines.

[1]Throughout the paper, "$\cdot$" between two vectors denotes the inner product; between scalars, it denotes ordinary multiplication.

3 BERNOULLI PREDICTION SETS (BPS)

While in CP, prediction sets are typically constructed by thresholding a score function (see (1)), we develop our approach within a more general framework of randomized prediction sets. In this section, we formalize randomized, parameterized prediction sets in full generality, and in Section 4 we focus on how to derive such prediction sets from a credal set predictor in an optimal manner.

For each input $\boldsymbol{x}_i$, a Bernoulli prediction set is specified by an *inclusion probability vector* $\boldsymbol{b}_i \in [0,1]^K$, where b_{ik} denotes the probability that class k is included in the predicted set. The realized (random) prediction set is obtained by sampling independent Bernoulli variables

$$z_{ik} \sim \mathrm{Bernoulli}(b_{ik}), \quad \mathcal{C}_{\text{BPS}}(\boldsymbol{x}_i) := \{k \in [K] : z_{ik} = 1\}. \tag{5}$$

Standard (deterministic) prediction sets are a special case of BPS with $\boldsymbol{b}_i \in \{0,1\}^K$. With this probabilistic formulation, we redefine both the set size and the conditional coverage in expectation. For input $\boldsymbol{x}_i$, the (expected) set size is

$$\mathbb{E}\big[|\mathcal{C}_{\text{BPS}}(\boldsymbol{x}_i)|\big] = \sum_{k=1}^{K}\mathrm{P}(z_{ik}=1) = \sum_{k=1}^{K} b_{ik} = \boldsymbol{b}_i \cdot \mathbf{1}, \tag{6}$$

and the conditional coverage is

$$\mathrm{P}\big(y_i \in \mathcal{C}_{\text{BPS}}(\boldsymbol{x}_i) \mid \boldsymbol{x}_i\big) = \sum_{k=1}^{K} p_{ik}\,\mathrm{P}(z_{ik}=1) = \boldsymbol{b}_i \cdot \boldsymbol{p}_i. \tag{7}$$

Note that in both measures $\boldsymbol{b}_i$ is assumed to be fixed; i.e., derived deterministically as a function of $\boldsymbol{x}_i$. Therefore, the probability is over the randomness of the sets being sampled from $\boldsymbol{b}_i$, and the label y being sampled from $\boldsymbol{p}_i$. In some cases, the $\boldsymbol{b}_i$ itself will be the outcome of a process that is random (e.g., over the randomness of the calibration set in case some conformalization is used), and in that case, the metrics will be evaluated in expectation over $\boldsymbol{b}_i$ as well.

Reformulating APS as BPS. Just to illustrate the generality of the definition, we show that adaptive prediction sets (APS), for any value of the conformal threshold, are a special case of BPS. Specifically, given a label distribution $\boldsymbol{\pi}_i$, let σ_i be a permutation of $[K] := \{1, \ldots, K\}$ that orders the labels in decreasing order of probability, i.e.,

$$\tilde{\pi}_{ij} := \pi_{i,\sigma_i(j)}, \quad \text{with} \quad \tilde{\pi}_{i1} \geq \tilde{\pi}_{i2} \geq \cdots \geq \tilde{\pi}_{iK}. \tag{8}$$

For any threshold $\tau \in [0,1]$, APS defines

$$L(\boldsymbol{\pi}_i, \tau) = \min\left\{k \in [K] : \sum_{j=1}^{k}\tilde{\pi}_{ij} \geq \tau\right\}$$

and constructs a randomized set as

$$\mathcal{C}_{\text{APS}}(\boldsymbol{x}_i, \boldsymbol{\pi}_i, \tau, u) = \begin{cases} \text{top } L(\boldsymbol{\pi}_i, \tau) - 1 \text{ labels}, & \text{if } u \leq u_0 \\ \text{top } L(\boldsymbol{\pi}_i, \tau) \text{ labels}, & \text{otherwise} \end{cases} \tag{9}$$

with

$$u \sim \text{Uniform}[0,1] \text{ and } u_0 = \tilde{\pi}^{-1}_{iL(\boldsymbol{\pi}_i,\tau)} \left[\sum_{j=1}^{L(\boldsymbol{\pi}_i,\tau)} \tilde{\pi}_{ij} - \tau \right].$$

In the BPS formulation, this construction corresponds to an inclusion probability vector defined over the original class indices as

$$b_{i,\sigma_i(j)} = \begin{cases} 1, & j < L(\boldsymbol{\pi}_i, \tau) \\ 1 - u_0, & j = L(\boldsymbol{\pi}_i, \tau) \\ 0, & j > L(\boldsymbol{\pi}_i, \tau) \end{cases}. \tag{10}$$

Thus, APS is a Bernoulli prediction set in which all labels with rank strictly below the threshold $L(\boldsymbol{\pi}_i, \tau)$ are deterministically included, the boundary label is included with probability $1 - u_0$, and all lower-ranked labels are excluded. While APS aims for a specific objective under a specific constraint (which is the minimum set size while preserving conditional coverage over a single predictive distribution), BPS is a general framework over which we can define more general constraints and objectives, as we further discuss.

4 SET PREDICTION FROM CREDAL SETS

We aim to construct prediction sets given a credal set predictor $\mathcal{Q}_i := \mathcal{Q}(\boldsymbol{x}_i) \subseteq \triangle^K$. A common representation of credal sets is the convex hull of m label distributions (the vertices of a convex polytope), i.e., $\mathcal{Q}_i = \text{Convex}(\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m)$. This representation aligns well with practice, as many credal set predictors are obtained from ensembles of probabilistic classifiers [Wang et al., 2024, 2025, Nguyen et al., 2025]. While we develop our framework for this polytope representation, more general credal sets can also be handled. In particular, if a credal set is not already given as the convex hull of finitely many points, it can first be approximated by a (finite) enclosing polytope, and the vertices of this polytope can then be used within our framework.

There is no unique way to design prediction sets, regardless of whether the underlying predictor is a probabilistic classifier or a credal set predictor. Therefore, beyond marginal coverage, which is satisfied by any CP-based set predictor, we require additional desiderata and an explicit objective to guide the construction. For instance, achieving conditional coverage with the smallest possible set size, when given access to the true label distributions, is the principle underlying the design of APS. We aim to follow a design principle analogous to that of APS but tailored to credal set predictors. To this end, we replace the assumption of access to the true label distribution with the assumption of access to *valid* credal sets, defined as follows.

Definition 4.1 (valid credal sets)**.** A credal set $\mathcal{Q}_i$ as a prediction for $\boldsymbol{x}_i$ is said to be valid if $\boldsymbol{p}_i \in \mathcal{Q}_i$, that is, if the oracle label distribution lies within the credal set.

It is worth noting that assuming access to valid credal sets is a weaker assumption than having access to oracle label distributions. Indeed, the former still accounts for epistemic uncertainty and does not require precise knowledge of the ground truth. In the following, we first consider the setting in which this validity assumption holds for every input $\boldsymbol{x}_i$ and formulate our desiderata accordingly, leading to a concrete set predictor (Section 4.1). We then study the case in which this assumption holds only with high probability over random draws of data points (Section 4.2), and finally, the case where the validity of the credal sets is unknown (Section 4.3).

4.1 CASE I: VALID CREDAL SETS

We begin with the setting in which the credal set predictor outputs a valid credal set for every input, that is, $\forall \boldsymbol{x}_i,\ \boldsymbol{p}_i \in \mathcal{Q}_i$. Given a valid credal set $\mathcal{Q}_i$, our objective is to construct a prediction set $\mathcal{C}(\boldsymbol{x}_i)$ based solely on $\mathcal{Q}_i$. Since the true label distribution lies in the credal set, a natural desideratum is that the resulting prediction set achieves conditional coverage of a nominal level $1 - \alpha$. However, because the exact distribution $\boldsymbol{p}_i$ is unknown, this requirement must hold uniformly over all distributions contained in $\mathcal{Q}_i$.

Desideratum 1 (Conditional Coverage). For valid credal sets, the prediction sets should satisfy conditional coverage uniformly over the credal set — formally,

$$\forall \boldsymbol{p} \in \mathcal{Q}_i : \quad \mathrm{P}_{y \sim \text{Categorical}(\boldsymbol{p})} \left[y \in \mathcal{C}(\boldsymbol{x}_i) \right] \geq 1 - \alpha. \tag{11}$$

As mentioned before, credal set predictors, unlike probabilistic classifiers, are able to represent epistemic uncertainty. Prediction sets are meant to reflect the predictive uncertainty of the learner. Accordingly, a second natural desideratum is that the prediction sets deflate as epistemic uncertainty decreases. Although it is not straightforward to define a total order over arbitrary credal sets in terms of the amount of epistemic uncertainty they represent, this requirement can still be formalized under a partial order (subset) relationship: if one credal set is a subset of another, it reflects less epistemic uncertainty, and the resulting prediction set should be smaller or equal in size. Formally, the following desideratum should be achieved by any prediction set that is adaptive to epistemic uncertainty.

Desideratum 2 (Epistemic Adaptivity). Given two valid credal sets $\mathcal{Q}_i$ and $\mathcal{Q}'_i$ at point $\boldsymbol{x}_i$ such that $\boldsymbol{p}_i \in \mathcal{Q}_i \subseteq \mathcal{Q}'_i$,

the prediction set constructed using $\mathcal{Q}_i$ ought to be smaller or equal in size to that constructed using $\mathcal{Q}'_i$.

We design a set predictor that satisfies the above desiderata while minimizing the set size as our objective. This leads to the design of the optimal Bernoulli prediction sets.

Optimal Bernoulli Prediction Sets. Over the previously defined BPS framework (Section 3), we aim to find $\boldsymbol{b}_i$'s with the objective of finding the smallest possible (random) prediction set that satisfies both desiderata. Formally, for a given credal set $\mathcal{Q}_i = \mathrm{Convex}(\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m)$, we determine an inclusion probability vector $\boldsymbol{b}_i^\lambda$ as a solution to the following optimization problem:

$$\boldsymbol{b}_i^\lambda = \underset{\boldsymbol{b}\in[0,1]^K}{\arg\min}\, \boldsymbol{b}\cdot\mathbf{1} \quad \text{s.t.} \quad \forall j \in [m]: \quad \boldsymbol{b}\cdot\boldsymbol{\pi}_i^{(j)} \geq \lambda, \tag{12}$$

where λ is the threshold determining the required level of conditional coverage. In particular, to achieve $1-\alpha$ conditional coverage, we set $\lambda = 1-\alpha$ and refer to the resulting predictor as $\mathrm{BPS}(1-\alpha)$.

It turns out that the optimization problem (12) is a linear program and can therefore be solved efficiently using standard solvers. In fact, it can be viewed as a multi-dimensional fractional knapsack problem. As a consequence of this linear structure, the solution $\boldsymbol{b}_i^\lambda$ contains at most $\min(K, m)$ fractional components, i.e., entries $b_{ik}^\lambda \in (0,1)$.

In the following, we show that the resulting prediction set $\mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_i)$ is the smallest set (in expectation) that conforms with the predefined desiderata when given access to a valid credal set.

Proposition 4.2. *For any $\alpha \in (0,1)$ and a given credal set $\mathcal{Q}_i = \mathrm{Convex}(\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m)$, the prediction set $\mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_i)$ in (5), constructed using $\mathbf{b}_i^{1-\alpha}$ from (12), is the smallest (possibly randomized) prediction set that satisfies $(1-\alpha)$ conditional coverage with respect to every distribution $\boldsymbol{p} \in \mathcal{Q}_i$.*

All proofs are deferred to Appendix B. Moreover, Desideratum 2 is satisfied by construction: the smallest prediction set that guarantees conditional coverage for a given credal set does not expand as epistemic uncertainty decreases. Indeed, a smaller credal set induces fewer constraints in (12), and therefore, the optimal (expected) set size cannot increase (and may decrease).

Proposition 4.3. *Consider two valid credal sets $\mathcal{Q}_i$ and $\mathcal{Q}'_i$ at point $\boldsymbol{x}_i$ such that $\boldsymbol{p}_i \in \mathcal{Q}_i \subseteq \mathcal{Q}'_i$. Let $\boldsymbol{b}_i^\lambda$ and $\boldsymbol{b}_i'^\lambda$ be the solutions to (12) corresponding to $\mathcal{Q}_i$ and $\mathcal{Q}'_i$, respectively. Then we have $\boldsymbol{b}_i^\lambda \cdot \mathbf{1} \leq \boldsymbol{b}_i'^\lambda \cdot \mathbf{1}$.*

In Appendix E, we provide toy examples that illustrate this property in a clear and intuitive manner. Another desirable property of the set predictor in Equation (12) is that it recovers APS when the credal set is a singleton.

Proposition 4.4. *For a fixed threshold λ and a singleton credal set $\mathcal{Q}_i = \{\boldsymbol{\pi}_i\}$, let $\boldsymbol{b}_i^\lambda$ be the solution of (12) with the single constraint given by $\boldsymbol{\pi}_i$. Then $\boldsymbol{b}_i^\lambda$ is equal to the Bernoulli inclusion vector of APS defined in (10) for the same $\boldsymbol{\pi}_i$ and the threshold λ. Consequently, the resulting prediction sets constructed by BPS and APS are equivalent in expectation.*

4.2 CASE II: PARTIALLY VALID CREDAL SETS

We now consider the case where credal sets are not guaranteed to be valid for every input, but instead are valid with high probability — formally,

$$\mathrm{P}_{\boldsymbol{x}_i}[\boldsymbol{p}_i \in \mathcal{Q}_i] \geq 1-\epsilon\,. \tag{13}$$

From a frequentist perspective, this means that over a random draw of inputs, the associated credal sets are valid for at least a $1-\epsilon$ fraction of cases. Such partially valid credal sets may arise, for instance, from conformalized credal set predictors [Javanmardi et al., 2024], where the credal sets satisfy a marginal CP guarantee of containing the label distribution with high probability.

In this setting, we again apply $\mathrm{BPS}(1-\alpha)$ directly to $\mathcal{Q}_i$, without using additional CP calibration. As a consequence, the resulting prediction sets satisfy the following PAC-style conditional coverage guarantee:

$$\mathrm{P}_{\boldsymbol{x}_i}\left[\boldsymbol{b}_i^{1-\alpha}\cdot\boldsymbol{p}_i \geq 1-\alpha\right] \geq 1-\epsilon. \tag{14}$$

Intuitively, on at least a $(1-\epsilon)$ fraction of inputs, namely, those for which the credal sets are valid, the conditional coverage is at least at the desired $(1-\alpha)$ level.

Proposition 4.5. *Suppose a credal set predictor outputs sets $\mathcal{Q}_i = \mathrm{Convex}(\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m)$ satisfying*

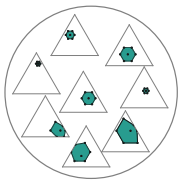
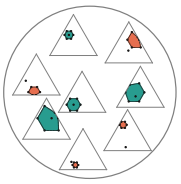
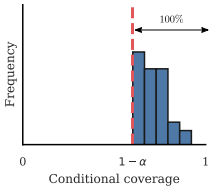
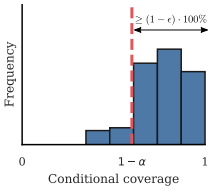
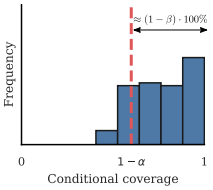
$$\mathrm{P}_{\boldsymbol{x}_i}\left[\boldsymbol{p}_i \in \mathcal{Q}_i\right] \geq 1-\epsilon.$$

For any $\alpha \in (0,1)$, let $\mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_i)$ be constructed as in (5) using the solution $\boldsymbol{b}_i^{1-\alpha}$ of (12). Then

$$\mathrm{P}_{\boldsymbol{x}_i}\left[\mathrm{P}\left(y_i \in \mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_i) \mid \boldsymbol{x}_i\right) \geq 1-\alpha\right] \geq 1-\epsilon.$$

It is worth noting that the validity of the credal set $\mathcal{Q}_i$ is a sufficient condition for the prediction set $\boldsymbol{b}_i^\lambda$ in (12) to guarantee λ conditional coverage with respect to the true label distribution $\boldsymbol{p}_i$, but it is not necessary. In particular, even if $\boldsymbol{p}_i \notin \mathcal{Q}_i$, the resulting Bernoulli vector may still satisfy $\boldsymbol{b}_i^\lambda \cdot \boldsymbol{p}_i \geq \lambda$. That being said, the guarantee in (14) should typically be interpreted as a worst-case statement. In practice, the realized conditional coverage with respect to $\boldsymbol{p}_i$ may hold for a substantially larger fraction than $1-\epsilon$.

Table 1: Methods and guarantees across the three credal set settings.

	Valid Credal Sets	**Partially Valid Credal Sets**	**Credal Sets with Unknown Validity**
			
	$\boldsymbol{p}_i \in \mathcal{Q}_i \;\; \forall \boldsymbol{x}_i$	$\mathrm{P}_{\boldsymbol{x}_i}[\boldsymbol{p}_i \in \mathcal{Q}_i] \geq 1-\epsilon$	unknown validity
Set Predictor	$\mathrm{BPS}(1-\alpha)$	$\mathrm{BPS}(1-\alpha)$	Calibrated $\mathrm{BPS}(\lambda^\star)$
Guarantee Type			
	$\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{p}_i \geq 1-\alpha \;\; \forall \boldsymbol{x}_i$	$\mathrm{P}_{\boldsymbol{x}_i}\left[\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{p}_i \geq 1-\alpha\right] \geq 1-\epsilon$	$\mathrm{P}_{\mathcal{D}_\dagger}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{p}_{n+1} \geq 1-\alpha\right] \geq 1-\beta$

So far, in both Case I (valid credal sets) and Case II (partially valid credal sets), the prediction sets are obtained directly by solving the optimization problem in (12). In these two settings, $\mathrm{BPS}(1-\alpha)$ yields prediction sets that either satisfy conditional coverage for every input (Case I) or satisfy it with high probability (Case II), without requiring an additional conformal calibration step.

4.3 CASE III: CREDAL SETS WITH UNKNOWN VALIDITY

We finally consider the most general setting, in which no validity assumption is imposed on the outputs of the credal set predictor. In this case, the optimization (12) alone does not provide any formal guarantee on the resulting conditional coverage. Therefore, in order to obtain a distribution-free guarantee, we resort to a conformal framework. Since the predictor in (12) is monotone in λ (i.e., the resulting sets inflate as λ increases), we adopt the conformal risk control framework, which calibrates λ to control a user-defined notion of risk.

As in the previous two cases, our goal is to obtain a conditional-type guarantee with respect to the true label distribution $\boldsymbol{p}_i$. To this end, we consider access to a holdout first-order calibration data set $\mathcal{D}_{\mathrm{cal}}^{\mathrm{first}} = \{(\boldsymbol{x}_i, \boldsymbol{p}_i)\}_{i=1}^n$ that is exchangeable with the future test point $(\boldsymbol{x}_{n+1}, \boldsymbol{p}_{n+1})$. We define the risk as the indicator that the conditional coverage falls below the target level $1-\alpha$, i.e.,

$$\mathcal{L}(\boldsymbol{x}_i, \lambda) := \mathbb{1}\left[\boldsymbol{b}_i^\lambda \cdot \boldsymbol{p}_i < 1-\alpha\right]. \tag{15}$$

For the set predictor in (12), the calibration data $\mathcal{D}_{\mathrm{cal}}^{\mathrm{first}}$, and the risk defined in (15), CRC determines the threshold $\lambda^\star$ as the solution to

$$\inf\left\{\lambda : \sum_{i=1}^{n} \mathbb{1}\left[\boldsymbol{b}_i^\lambda \cdot \boldsymbol{p}_i \geq 1-\alpha\right] \geq \lceil (1-\beta)(n+1) \rceil\right\}, \tag{16}$$

which bounds the expected risk at the user-specified tolerance $\beta \in [0,1]$. This translates into the guarantee that

$$\mathrm{P}_{\mathcal{D}_\dagger}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{p}_{n+1} \geq 1-\alpha\right] \geq 1-\beta, \tag{17}$$

where $\mathcal{D}_\dagger = \mathcal{D}_{\mathrm{cal}}^{\mathrm{first}} \cup \{(\boldsymbol{x}_{n+1}, \boldsymbol{p}_{n+1})\}$. In our experiments, we refer to this approach as calibrated BPS with first-order data. Table 1 provides an overview of all three cases, the corresponding set predictors, and the guarantees they deliver.

Proposition 4.6. *Let a credal set predictor output $\mathcal{Q}_i = \mathrm{Convex}(\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m)$ without any validity assumption, and let $\alpha, \beta \in (0,1)$ be arbitrary. Let $\lambda^\star$ be defined as in (16) using the first-order calibration data $\mathcal{D}_{\mathrm{cal}}^{\mathrm{first}}$ that is exchangeable with the test point. Let $\mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_i)$ be constructed as in (5) using $\boldsymbol{b}_i^{\lambda^\star}$ from (12). Then the resulting prediction sets satisfy*

$$\mathrm{P}_{\mathcal{D}_\dagger}\left[\mathrm{P}\left(y_{n+1} \in \mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_{n+1}) \mid \boldsymbol{x}_{n+1}\right) \geq 1-\alpha\right] \geq 1-\beta.$$

This is again a PAC-style guarantee, similar to the one obtained in the case of partially valid credal sets. However, in contrast to that setting, the parameter β is user specified and directly controls the tolerance for violations of the conditional coverage requirement. Moreover, since this is a conformal guarantee, it is essentially tight in finite samples: under exchangeability, the coverage probability is also upper bounded by $1-\beta+\frac{2}{n+1}$, reflecting the usual sharpness of conformal bounds. Appendix C shows how these conditional guarantees translate into marginal ones.

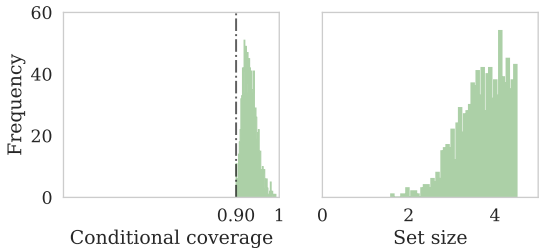


Figure 1: Histograms of the conditional coverage and (expected) set size of BPS$(1-\alpha)$ under valid credal sets. Here, $1-\alpha = 0.9$.

Remark (calibration beyond BPS). Although we instantiated the calibration mechanism with BPS, the conformalization step is not restricted to this particular set predictor. In principle, for any set predictor $\mathcal{C}_\lambda(\boldsymbol{x}_i)$ parameterized by λ, one may define the risk as the indicator that the conditional coverage falls below $1-\alpha$. As long as this risk is non-increasing in λ, conformal risk control can be applied to determine a calibrated $\lambda^\star$ that yields the same distribution-free guarantee (see Appendix I).

Remark (data requirements and the zero-order proxy). The first-order data used in this case are required only for calibration and not for training. Credal set predictors are typically trained using standard zero-order data, and therefore, the calibration set does not need to be very large. However, when no first-order calibration data (or reliable approximations thereof) are available, one possible approach is to use the one-hot encoding $\boldsymbol{e}_{y_i}$ of the observed labels as a proxy for $\boldsymbol{p}_i$ and treat these as first-order observations within the same calibration procedure. This yields an approximate version of the guarantee (see Appendix H). In our experiments, we refer to this variant as *calibrated BPS with zero-order data*, in contrast to the original *calibrated BPS with first-order data*.

5 EXPERIMENTS

In this section, we evaluate the performance of our proposed BPS under different scenarios. First, we consider a setting in which valid credal sets are available. Since such a setup does not yet exist in practice, we simulate it synthetically. Next, we study a scenario in which the credal sets are only partially valid. In these two cases, we simply apply BPS$(1-\alpha)$ without using conformal prediction. Finally, we turn to the setting where the validity of the credal sets is unknown. In this case, we apply BPS together with the proposed calibration procedure. All implementations and experiments can be found on our GitHub repository. [2]

[2]The link to the code: `https://github.com/alireza-javanmardi/conformal-BPS`

Table 2: Performance of BPS$(1-\alpha)$ on the partially valid credal sets for CIFAR-10. Here $1-\alpha = 0.9$.

ϵ	Credal Cvg.	Cond. Sat.	Marg. Cvg.	Set Size
0.10	0.90 ± 0.01	0.97 ± 0.00	1.00 ± 0.00	6.55 ± 0.23
0.20	0.80 ± 0.01	0.93 ± 0.01	0.99 ± 0.00	1.72 ± 0.43
0.30	0.70 ± 0.01	0.84 ± 0.01	0.95 ± 0.01	1.18 ± 0.02

Datasets. For the partially valid and unknown-validity settings, we conduct experiments on three real-world datasets for which first-order data is available. These include *CIFAR-10* [Krizhevsky et al., 2009], a 10-class image classification dataset, for which *CIFAR-10H* [Peterson et al., 2019] provides label distributions over the test set; *ChaosNLI* [Nie et al., 2020], a 3-class natural language inference dataset; and *QualityMRI* [Obuchowicz et al., 2020, Schmarje et al., 2022], a two-class medical image classification dataset. For all these datasets, we additionally use the corresponding zero-order data for evaluation purposes. Details regarding these datasets, the train–test–calibration splits, and information about the models (both the credal set predictors and the probabilistic classifiers) are provided in Appendix F.

Metrics. In all experiments, we consider three evaluation metrics. Since the main focus of the paper is conditional coverage, we report the conditional coverage satisfaction (**Cond. Sat.**), defined as the complement of the risk in (15). We also report the (expected) **set size**, as defined in (6), and the (expected) coverage of the realized label (**Marg. Cvg.**). All metrics are averaged over the test inputs.

5.1 VALID CREDAL SETS

To illustrate the performance of BPS on valid credal sets, we first generate $n = 1000$ instances, each represented by a convex hull of $m = 10$ randomly selected label distributions with $K = 5$ classes. For each instance, we sample a random convex combination of the m distributions and treat it as the true label distribution. We then apply BPS$(1-\alpha)$ to each credal set and evaluate the conditional coverage and the set size accordingly. Figure 1 shows that the resulting prediction sets satisfy the conditional coverage property for every instance.

5.2 PARTIALLY VALID CREDAL SETS

In order to obtain credal sets that are partially valid, we consider conformalized credal set predictors [Javanmardi et al., 2024]. To that end, we use a probabilistic classifier as the base learner together with a set of first-order calibration data $\mathcal{D}_{\text{cal}}^{\text{first}} = \{(\boldsymbol{x}_i, \boldsymbol{p}_i)\}_{i=1}^n$. On this data, we compute the total variation distance between the true label distribution $\boldsymbol{p}_i$ and the estimated one $\boldsymbol{\pi}_i$, forming the score set $\mathcal{S}_{\text{TV}} = \{\text{TV}(\boldsymbol{\pi}_i, \boldsymbol{p}_i)\}_{i=1}^n$. For any future test input $\boldsymbol{x}_{n+1}$,

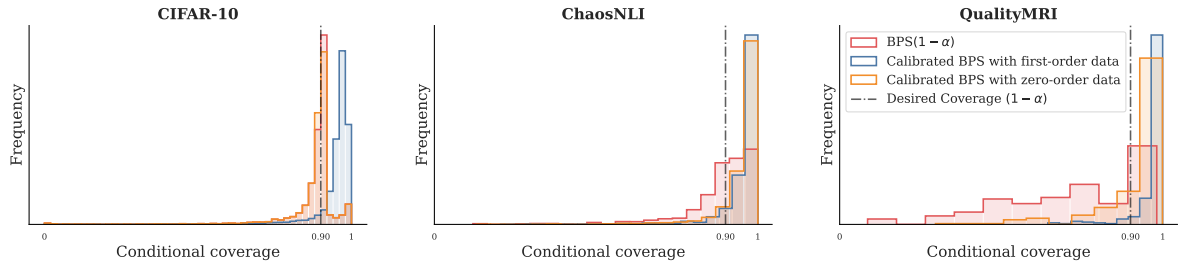


Figure 2: Conditional coverage histograms of BPS$(1-\alpha)$ and calibrated BPS with zero- and first-order data on credal sets with unknown validity across three real-world datasets. $1-\alpha = 0.9$ and $1-\beta = 0.9$.

Table 3: Performance comparison of BPS$(1-\alpha)$ and calibrated BPS with zero- and first-order data on credal sets with unknown validity across three real-world datasets. $1-\alpha = 0.9$ and $1-\beta = 0.9$.

	CIFAR-10			**ChaosNLI**			**QualityMRI**		
Approach	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size
BPS $(1-\alpha)$	0.52 ± 0.00	0.90 ± 0.00	1.09 ± 0.00	0.61 ± 0.01	0.89 ± 0.01	2.00 ± 0.02	0.33 ± 0.05	0.81 ± 0.02	1.36 ± 0.04
Calibrated BPS with first-order	0.90 ± 0.01	0.98 ± 0.00	1.45 ± 0.05	0.91 ± 0.02	0.95 ± 0.01	2.40 ± 0.04	0.94 ± 0.04	0.99 ± 0.01	1.93 ± 0.02
Calibrated BPS with zero-order	0.52 ± 0.00	0.90 ± 0.00	1.09 ± 0.00	0.90 ± 0.01	0.95 ± 0.01	2.35 ± 0.04	0.77 ± 0.10	0.96 ± 0.03	1.81 ± 0.07

the conformalized credal set is then defined as

$$\mathcal{Q}_{n+1} = \{\boldsymbol{q} \in \triangle^K : \mathrm{TV}(\boldsymbol{\pi}_{n+1}, \boldsymbol{q}) \leq \mathbb{Q}\,(1-\epsilon; \mathcal{S}_{\mathrm{TV}})\}.$$

For every $\epsilon \in (0,1)$, this credal set is guaranteed to cover the true label distribution with high probability, i.e.,

$$\mathrm{P}_{\mathcal{D}_\dagger}[\boldsymbol{p}_{n+1} \in \mathcal{Q}_{n+1}] \geq 1-\epsilon.$$

In Theorem D.1 in Appendix D, we provide an analytical derivation of the vertices of the credal sets constructed in this manner. Given the credal sets defined as the convex hull of their vertices, we then apply BPS$(1-\alpha)$ on the test set. For this method, we randomly split the evaluation data into calibration and test subsets, repeat this procedure 10 times with different random seeds, and report the results averaged over these runs.

In Table 2, we show the performance of this method on CIFAR-10 for $\epsilon \in \{0.1, 0.2, 0.3\}$. For this setting, we also report the coverage of the credal sets (**Credal Cvg.**), which matches the theoretical guarantee. As ϵ increases, the credal sets shrink in general, which results in smaller prediction sets. Furthermore, it can be seen that the conditional coverage satisfaction is in general higher than $1-\epsilon$, which is also expected, as explained in Section 4.2. This is mainly because prediction sets constructed from invalid credal sets may also satisfy conditional coverage. We have also applied the same approach to the other two datasets and provide the corresponding results in Table 4 in Appendix G.

5.3 CREDAL SETS WITH UNKNOWN VALIDITY

For this case, we consider the Credal Ensembling (CreEns) approach [Nguyen et al., 2025] as a credal set predictor, where the credal set is constructed as the convex hull of the softmax outputs of an ensemble of neural networks (see Appendix F.1 for further details on the CreEns model). For each dataset, we randomly set aside a portion of the data for calibration and apply calibrated BPS with both zero- and first-order data. This random split into calibration and test sets is repeated 10 times with different random seeds; prediction sets are constructed in each run, and the reported results are averaged over these seeds. In addition, we consider BPS$(1-\alpha)$ (i.e., BPS without calibration at nominal level $1-\alpha$) as a baseline.

Figure 2 compares the conditional coverage over the test set for the three approaches on each dataset, and Table 3 provides the full performance comparison over the test sets for all three approaches and datasets. As expected, BPS calibrated with first-order data performs best, satisfying the desired conditional coverage for at least a $1-\beta = 0.9$ fraction of cases. In contrast, BPS$(1-\alpha)$ violates conditional satisfaction in all cases, although it produces the most efficient (smallest) prediction sets. Calibrated BPS with zero-order data improves over BPS$(1-\alpha)$ in terms of conditional coverage; in some cases, such as ChaosNLI, it performs comparably to BPS calibrated with first-order data. Overall, it lies between the other two methods with respect to both efficiency and conditional coverage.

The results in all three cases support our theoretical findings. In Appendix H and Appendix I, we further investigate the effect of alternative calibration strategies—corresponding to different risk definitions—on these three datasets in the setting of credal sets with unknown validity. Moreover, since the guarantee in (17) can be achieved by any set predictor and is not restricted to BPS, we additionally examine the performance of the proposed calibration strategy when applied to alternative set predictors. Appendix J additionally evaluates all methods under distribution shift.

6 CONCLUSION

In this paper, we consider the problem of set prediction in a classification setting where the underlying model is a credal set predictor that outputs a set of label distributions for each input. We start with the case where the credal sets are valid, meaning the true label distribution lies within the credal set for every input. We propose a parametric set predictor, the Bernoulli prediction set (BPS), that outputs the smallest possible set guaranteeing conditional coverage at each input. When validity holds only with high probability over inputs, BPS can be applied to achieve a PAC-style conditional coverage guarantee; that is, with high probability, the achieved conditional coverage is at least the desired level. When the validity of the credal sets is unknown, BPS cannot provide guarantees on its own. For this case, we use first-order calibration data together with a risk function, defined as the indicator that conditional coverage is below the desired level, to calibrate BPS through the Conformal Risk Control (CRC) framework. This approach also leads to a PAC-style conditional coverage guarantee. In this case, the amount of violation of conditional coverage can be controlled by the user-specified error rate of CRC.

Acknowledgements

Alireza Javanmardi was supported by the Klaus Tschira Stiftung (project 00.019.2024). Willem Waegeman was supported by the Flemish Government under the Flanders AI Research program. The authors are thankful to Stefan Heid for his valuable insights on Theorem D.1.

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*, 2023.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, 2024a.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2024b. URL https://arxiv.org/abs/2411.11824.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Ilia Azizi, Juraj Bodik, Jakob Heiss, and Bin Yu. CLEAR: Calibrated learning for epistemic and aleatoric risk. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RY4IHaDLik.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2019.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*. PMLR, 2015.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015.

Luben Cabezas, Vagner S Santos, Thiago R Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. *arXiv preprint arXiv:2502.06995*, 2025.

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*, 2024a.

Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=AH5KwUSsln.

Michele Caprio, David Stutz, Shuo Li, and Arnaud Doucet. Conformalized credal regions for classification with ambiguous ground truth. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=L7sQ8CW2FY.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Paul Hofman, Timo Löhr, Maximilian Muschalik, Yusuf Sale, and Eyke Hüllermeier. Efficient credal prediction through decalibration. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BqOmsYIe7M.

Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54, 1996.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 2021.

Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=vBah12uVbD.

Hamed Karimi and Reza Samavi. Evidential uncertainty sets in deep classifiers using conformal prediction. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, 2024. URL https://proceedings.mlr.press/v230/karimi24a.html.

Shayan Kiyani, George J. Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. Crowdsourcing in computer vision. *Foundations and Trends® in computer graphics and Vision*, 10, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

Timo Löhr, Paul Hofman, Felix Mohr, and Eyke Hüllermeier. Credal prediction based on relative likelihood. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rKM3oqruN3.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.

Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Credal ensembling in multi-class classification. *Machine Learning*, 114(1), 2025.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

Rafal Obuchowicz, Mariusz Oszust, and Adam Piorkowski. Interobserver variability in quality assessment of magnetic resonance images. *BMC Medical Imaging*, 2020.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33, 2020.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett. Integrating uncertainty awareness into conformalized quantile regression. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? - a data-centric image classification benchmark for noisy and ambiguous label estimation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Nature, 2022.

Peter Walley. *Statistical reasoning with imprecise probabilities*, volume 42. Springer, 1991.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, Hans Hallez, et al. Credal deep ensembles for uncertainty quantification. *Advances in Neural Information Processing Systems*, 2024.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cv2iMNWCsh.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 2002.

Appendix A: Related Work

Set prediction under epistemic uncertainty. The connection between conformal prediction and epistemic uncertainty-aware predictors has recently attracted increasing attention. Rossellini et al. [2024] incorporate epistemic uncertainty into conformalized quantile regression to improve conditional coverage, while Azizi et al. [2026] construct prediction sets by adaptively balancing epistemic and aleatoric uncertainty. Our work differs from these two as we focus on classification, and their regression-based approaches are not directly transferable. Furthermore, these methods often rely on heuristics that do not provide the same theoretical guarantees as our framework. Karimi and Samavi [2024] introduce a nonconformity score based on uncertainty estimates derived from evidential models, and Cabezas et al. [2025] integrate epistemic uncertainty into conformal prediction by fitting a Bayesian model on top of nonconformity scores. What separates our work from these Bayesian and evidential approaches is that we study how epistemic uncertainty, represented explicitly via credal sets, can be incorporated into conformal prediction. Our focus is on providing finite-sample conditional coverage guarantees, whereas previous methods lack such formal properties in this setting. The problem of deriving prediction sets from credal sets has also been studied in the imprecise probability literature [Caprio et al., 2024a]. However, such approaches typically do not provide finite-sample coverage guarantees without additional structural assumptions.

Availability of first-order data. In practice, many datasets are labeled by human annotators, and in many cases, multiple annotators are involved. This approach is becoming increasingly common, with crowdsourcing tools becoming an integral component of dataset collection, making multiple annotations per data instance more accessible [Kovashka et al., 2016, Sorokin and Forsyth, 2008]. Disagreement among annotators is also very common, especially in natural language tasks [Abercrombie et al., 2023, Nie et al., 2020] and computer vision [Beyer et al., 2020, Peterson et al., 2019, Schmarje et al., 2022]. This facilitates the availability of such data in practice. In our case, training the credal set predictor relies solely on standard zero-order data; first-order data is only required during the calibration step and need not be large. The number of calibration samples primarily affects the sharpness of the coverage distribution in conformal prediction, and in practice, tens to a few hundred such samples are often sufficient.

Appendix B: Proofs

Proposition 4.2

Proof. We show that this set achieves at least $1-\alpha$ expected conditional coverage with respect to any $\boldsymbol{p} \in \mathcal{Q}_i$. Since $\boldsymbol{p}$ lies in the convex hull of $\{\boldsymbol{\pi}_i^{(j)}\}_{j=1}^m$, we can write $\boldsymbol{p} = \sum_{j=1}^m \eta_j \boldsymbol{\pi}_i^{(j)}$ for some $\eta_j \in [0,1]$ with $\sum_{j=1}^m \eta_j = 1$. It follows that $\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{p} = \sum_{j=1}^m \eta_j \, (\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{\pi}_i^{(j)})$, and by definition, $\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{\pi}_i^{(j)} \geq 1-\alpha$ for all j. Then

$$\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{p} = \sum_{j=1}^m \eta_j (\boldsymbol{b}_i^{1-\alpha} \cdot \boldsymbol{\pi}_i^{(j)}) \geq 1-\alpha \sum_{j=1}^m \eta_j = 1-\alpha.$$

The fact that it is the minimal set directly follows from the definition of Equation (12). □

Proposition 4.3

Proof. Since $\boldsymbol{p}_i \in \mathcal{Q}_i \subseteq \mathcal{Q}_i'$, the solution to the optimization in Equation (12) with $\mathcal{Q}_i'$ is also feasible for the same optimization with $\mathcal{Q}_i$. Therefore, $\boldsymbol{b}_i^{\lambda} \cdot \mathbf{1} \leq \boldsymbol{b}_i'^{\lambda} \cdot \mathbf{1}$. □

Proposition 4.4

Proof. For a singleton credal set $\mathcal{Q}_i = \{\boldsymbol{\pi}_i\}$, the optimization problem in Equation (12) reduces to

$$\boldsymbol{b}_i^{\lambda} = \arg\min_{\boldsymbol{b}} \; \boldsymbol{b} \cdot \mathbf{1} \quad \text{s.t.} \quad \boldsymbol{b} \cdot \boldsymbol{\pi}_i \geq \lambda.$$

This is a fractional knapsack problem. Its optimal solution is obtained by sorting $\boldsymbol{\pi}_i$ in decreasing order and allocating inclusion probabilities to the corresponding labels greedily. Let $\tilde{\boldsymbol{\pi}}_i$ denote the sorted version of $\boldsymbol{\pi}_i$ such that $\tilde{\pi}_{i1} \geq \cdots \geq$

$\tilde{\pi}_{iK}$, and let $L(\boldsymbol{\pi}_i, \lambda) = \min\{k \in [K] : \sum_{j=1}^{k} \tilde{\pi}_{ij} \geq \lambda\}$. Then the optimal vector in the sorted coordinate system is $(1, \ldots, 1, b^{\lambda}_{iL(\boldsymbol{\pi}_i,\lambda)}, 0, \ldots, 0)$, where the first $L(\boldsymbol{\pi}_i, \lambda) - 1$ entries are equal to 1, the entries after $L(\boldsymbol{\pi}_i, \lambda)$ are 0, and

$$b^{\lambda}_{iL(\boldsymbol{\pi}_i,\lambda)} = \frac{\lambda - \sum_{c=1}^{L(\boldsymbol{\pi}_i,\lambda)-1} \tilde{\pi}_{ic}}{\tilde{\pi}_{iL(\boldsymbol{\pi}_i,\lambda)}}.$$

Reordering back to the original label indexing yields $\boldsymbol{b}^{\lambda}_i$, which satisfies $\boldsymbol{b}^{\lambda}_i \cdot \boldsymbol{\pi}_i = \lambda$.

This coincides exactly with the Bernoulli inclusion vector underlying APS in (10) for the same $\boldsymbol{\pi}_i$ and threshold λ. In both cases, the top $L(\boldsymbol{\pi}_i, \lambda) - 1$ labels are included deterministically, the $L(\boldsymbol{\pi}_i, \lambda)$-th label is included with probability $b^{\lambda}_{iL(\boldsymbol{\pi}_i,\lambda)}$, and all remaining labels are excluded. Hence, the BPS and APS constructions induce the same inclusion probabilities and therefore yield equivalent prediction sets in expectation. □

Proposition 4.5

Proof. On the event $\boldsymbol{p}_i \in \mathcal{Q}_i$, validity of the credal set implies $\boldsymbol{b}^{1-\alpha}_i \cdot \boldsymbol{p}_i \geq 1 - \alpha$ by construction of BPS. Since this event occurs with probability at least $1 - \epsilon$, the result follows. □

Proposition 4.6

Proof. By construction, $\lambda^\star$ controls the empirical risk in (15). The result follows directly from the distribution-free guarantee of conformal risk control, which ensures that the expected risk of the test point is bounded by β under exchangeability. □

Appendix C: From Conditional to Marginal Coverage Guarantees

The conditional coverage guarantee directly implies a marginal coverage guarantee as well. In particular, if $\boldsymbol{b}^{1-\alpha}_i \cdot \boldsymbol{p}_i \geq 1 - \alpha, \quad \forall \boldsymbol{x}_i$, then automatically $\mathrm{P}\left[y_{n+1} \in \mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_{n+1})\right] \geq 1 - \alpha$, where the probability is over the test point and the randomization of the set. However, in the PAC-style conditional coverage setting, the guarantee transfers to a marginal coverage guarantee of at least $(1 - \delta)(1 - \alpha)$, where δ denotes the corresponding error rate ($\delta = \epsilon$ under partially valid credal sets in Section 4.2, and $\delta = \beta$ under credal sets with unknown validity in Section 4.3). We show this for Case III; the argument for Case II is identical, with the probability taken over the random draw of the input instead of $\mathcal{D}_\dagger$. Indeed,

$$\mathrm{P}_{\mathcal{D}_\dagger}\left[y_{n+1} \in \mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_{n+1})\right] = \mathbb{E}_{\mathcal{D}_\dagger}\left[\mathrm{P}(y_{n+1} \in \mathcal{C}_{\mathrm{BPS}}(\boldsymbol{x}_{n+1}) \mid \boldsymbol{x}_{n+1})\right] = \mathbb{E}_{\mathcal{D}_\dagger}\left[\boldsymbol{b}^{\lambda^\star}_{n+1} \cdot \boldsymbol{p}_{n+1}\right] \geq (1 - \beta)(1 - \alpha) + \underbrace{\beta \cdot 0}_{\text{worst case}}.$$

In practice, however, we observe that the achieved marginal coverage under the PAC-style conditional guarantee is often substantially higher than this worst-case bound.

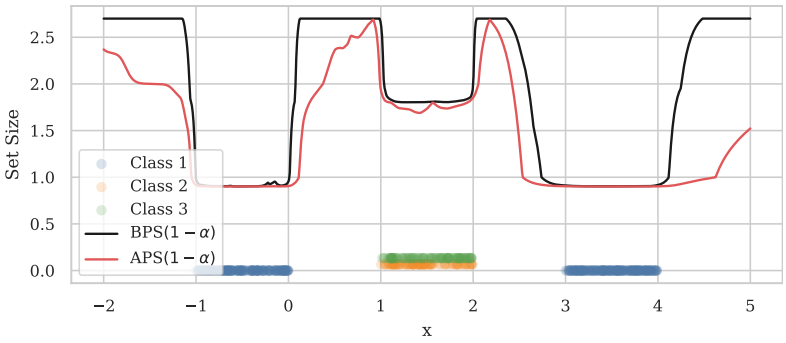
Appendix D: TV-Distance Neighborhood: A Specific Credal Set with Analytical Corner Points

In this section, we show that a credal set defined as the set of all label distributions in the simplex $\triangle^K$ with total variation distance less than or equal to d from $\boldsymbol{p}$ can be analytically described by a set of $K(K-1)$ corner points.

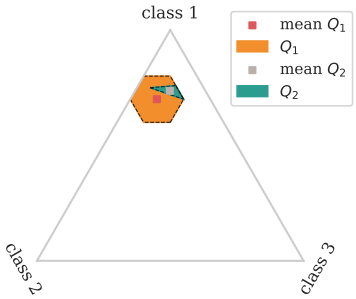
Theorem D.1. *The d-neighborhood of a distribution $\boldsymbol{p} \in \triangle^K$, i.e., the set of all distributions in $\triangle^K$ with a total variation distance $\leq d$, is a polytope defined by a finite number of corner points in $\triangle^K$.*

Proof. Let $\bar{\triangle}^K \supset \triangle^K$ denote the set of all $\boldsymbol{p} = (p_1, \ldots, p_K) \in \mathbb{R}^K$ such that $p_1 + \ldots + p_K = 1$ (i.e., also including negative entries). The set of all $\boldsymbol{q} \in \bar{\triangle}^K$ whose total variation distance from $\boldsymbol{p}$ is bounded by d, i.e.,

$$N_d(\boldsymbol{p}) := \left\{ \boldsymbol{q} \in \bar{\triangle}^K \mid \max_{A \subseteq [K]} \left| \sum_{k \in A} p_k - \sum_{k \in A} q_k \right| \leq d \right\},$$



(a) Example 1



(b) Example 2

Figure 3: Adaptivity of BPS versus APS to epistemic uncertainty. For both examples, we set $1 - \alpha = 0.9$.

is given by the convex polytope C with corner points $\boldsymbol{p}^{i,j} = \boldsymbol{p} + \boldsymbol{h}^{i,j}$, $1 \leq i \neq j \leq K$, where the i^{th} entry in $\boldsymbol{h}^{i,j}$ is $+d$, the j^{th} entry is $-d$, and all other entries are 0. For the direction $C \subset N_d(\boldsymbol{p})$, note that any $\boldsymbol{p}^{i,j}$ is obviously in $N_d(\boldsymbol{p})$. Moreover, for $\boldsymbol{q} = \sum_{i,j} \alpha_{i,j}\, \boldsymbol{p}^{i,j}$, where $\alpha_{i,j} \geq 0$ and $\sum_{i,j} \alpha_{i,j} = 1$, and any $A \subseteq [K]$,

$$\begin{aligned}
\sum_{k \in A} p_k - \sum_{k \in A} q_k &= \sum_{k \in A} p_k - \sum_{k \in A} \sum_{i,j} \alpha_{i,j}\, p_k^{i,j} \\
&= \sum_{k \in A} \sum_{i,j} \alpha_{i,j}\, (p_k - p_k^{i,j}) \\
&= \sum_{i,j} \alpha_{i,j} \sum_{k \in A} (p_k - p_k^{i,j}) \\
&\leq \max_{i,j} \sum_{k \in A} (p_k - p_k^{i,j}) \\
&\leq d.
\end{aligned}$$

To show the direction $N_d(\boldsymbol{p}) \subset C$, consider any $\boldsymbol{q} \in N_d(\boldsymbol{p})$ such that $\boldsymbol{q} \neq \boldsymbol{p}$. Let $A \subseteq [K]$ be a (smallest) subset that determines the total variation distance between $\boldsymbol{q}$ and $\boldsymbol{p}$. Then either $q_i > p_i$ for all $i \in A$, or $q_i < p_i$ for all $i \in A$. Consider the first case (without loss of generality) and let $(\cdot)$ be a permutation of $[K]$ such that $q_{(1)} - p_{(1)} \geq q_{(2)} - p_{(2)} \geq \ldots \geq q_{(K)} - p_{(K)}$. Then $A = \{(1), \ldots, (J)\}$ for some $1 \leq J < K$ and

$$\sum_{k \in [J]} q_{(k)} - p_{(k)} = d' \leq d\,.$$

We can then write $q_{(k)} = \sum_{i,j} \alpha_{i,j}\, p_{(k)}^{i,j}$ for suitably chosen $\alpha_{i,j} \geq 0$ with $\sum_{i,j} \alpha_{i,j} = 1$, which means that $\boldsymbol{q} \in C$. The $\alpha_{i,j}$ can be specified in a constructive way by processing the entries in $\boldsymbol{q}$ one by one, starting with the k for which the difference $|q_{(k)} - p_{(k)}|$ is smallest and proceeding to the larger ones.

Now, the d-neighborhood of $\boldsymbol{p}$ that we are looking for is given by the intersection $C \cap \Delta^K$, i.e., by the intersection of two convex polytopes. Thus, it is itself again a convex polytope. □

Appendix E: Toy Examples Illustrating Epistemic Adaptivity

Example 1. Consider the following toy example of a three-class classification problem, where $\boldsymbol{x}$ is a one-dimensional input. The data-generating mechanism is defined as follows:

- If $\boldsymbol{x}_i \in [-1, 0] \cup [3, 4]$, then y_i belongs to class 1.
- If $\boldsymbol{x}_i \in [1, 2]$, then y_i belongs to class 2 or 3 with equal probability, i.e., $\mathrm{P}(y_i = 2 \mid \boldsymbol{x}_i) = \mathrm{P}(y_i = 3 \mid \boldsymbol{x}_i) = 0.5$.

In other regions of the input space, namely $[-2, -1]$, $[0, 1]$, $[2, 3]$, and $[4, 5]$, no training data are observed. An ensemble of neural networks is trained on this data and treated as a credal set predictor. For inputs $\boldsymbol{x}_i \in [-2, 5]$, prediction sets

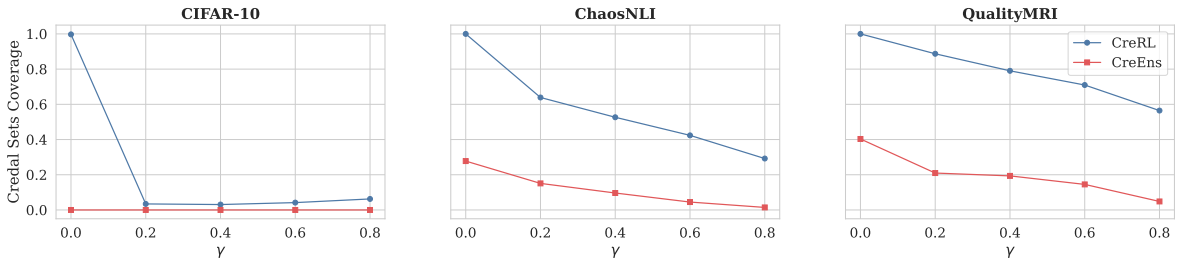


Figure 4: Credal sets coverage for different credal set predictors on the three real-world datasets.

are constructed using BPS$(1-\alpha)$. For comparison, we also aggregate the ensemble by taking the mean of its predictive distributions, treat it as a single probabilistic classifier, and construct prediction sets using APS$(1-\alpha)$. As shown in Figure 3a, in regions where sufficient training data are available, both set predictors behave similarly. However, in regions with no observed data, BPS reflects the increased epistemic uncertainty through larger prediction sets, whereas APS does not consistently account for this uncertainty.

Example 2. Consider two credal sets as illustrated in Figure 3b, where $\mathcal{Q}_2 \subset \mathcal{Q}_1$. As a consequence of Proposition 4.3 (epistemic adaptivity), the prediction set constructed by BPS for $\mathcal{Q}_1$ must be at least as large as the one constructed for $\mathcal{Q}_2$, since $\mathcal{Q}_1$ represents higher epistemic uncertainty. In this example, we also demonstrate that ignoring the credal set and instead working with the averaged predictor combined with APS does not satisfy this adaptivity property. Concretely, for $\mathcal{Q}_1$, the Bernoulli inclusion vector obtained by APS (applied to the mean predictor) is $[1, 1, 0]$, while BPS yields $[1, 0.91, 0.58]$. For $\mathcal{Q}_2$, APS produces $[1, 1, 0.23]$, whereas BPS results in $[1, 0.57, 0.71]$. The expected set size (i.e., the sum of inclusion probabilities) for BPS decreases from 2.49 to 2.28 when moving from $\mathcal{Q}_1$ to $\mathcal{Q}_2$, reflecting the reduction in epistemic uncertainty. In contrast, for APS, the expected set size increases from 2 to 2.23, violating epistemic adaptivity.

Appendix F: Experiments Details

F.1 MODELS

For the experiments on real-world datasets, we consider two credal set predictors: Credal Relative Likelihood (CreRL) [Löhr et al., 2025] and Credal Ensembling (CreEns) [Nguyen et al., 2025]. Both methods are based on an ensemble of neural networks. In all experiments, the number of ensemble members is fixed to $m = 20$. The CreRL framework considers a parameter γ (referred to as α in the original paper; to avoid confusion with our error rate α, we denote it by γ). It constructs the credal set from the m ensemble members, where the relative likelihood of the i^{th} member is limited to $\gamma + (i-1) \cdot \frac{1-\gamma}{m-1}, \forall i \in [m]$ by an early stopping strategy. The CreEns framework builds on standard ensemble training with m members. It first considers the mean predictor as a representative and computes the Euclidean distance of each of the m members to this representative. For a given γ, it then removes a γ fraction of the ensemble members with the largest Euclidean distance from the representative. All training details follow Löhr et al. [2025] and their GitHub repository[3]. In Figure 4, we plot the coverage of the credal sets, defined as the convex hull of the m predictors constructed by these methods, across three datasets for various values of γ. Note that the smaller the value of γ, the larger the resulting credal sets, hence the higher the coverage.

For the experiments in Section 5.2, since we required a probabilistic classifier, we considered the average predictor of CreRL ($\gamma = 1.0$) ensemble as the probabilistic classifier. We choose $\gamma = 1.0$ because in this setting the coverage of the credal set is not relevant; we only rely on the average predictor as a probabilistic classifier. For the experiments in Section 5.3, in order to balance credal coverage and set size, we used CreEns ($\gamma = 0.4$), resulting in an ensemble of size 12. Later, in Table 7, when reporting experiments with APS, we use the average predictor of this ensemble as the probabilistic classifier.

F.2 DATASETS

CIFAR-10. The CIFAR-10 dataset [Krizhevsky et al., 2009] is an image classification dataset with 10 classes, containing 50K training and 10K test instances. A variant of this dataset, namely CIFAR-10H [Peterson et al., 2019], provides multiple

[3]https://github.com/timoverse/credal-prediction-relative-likelihood

Table 4: Performance of BPS$(1-\alpha)$ on the partially valid credal sets for different datasets. Here $1-\alpha = 0.9$.

	CIFAR-10				ChaosNLI				QualityMRI			
ϵ	Credal Cvg.	Cond. Sat.	Marg. Cvg.	Set Size	Credal Cvg.	Cond. Sat.	Marg. Cvg.	Set Size	Credal Cvg.	Cond. Sat.	Marg. Cvg.	Set Size
0.10	0.90 ± 0.01	0.97 ± 0.00	1.00 ± 0.00	6.55 ± 0.23	0.89 ± 0.02	0.98 ± 0.01	0.92 ± 0.00	2.69 ± 0.00	0.90 ± 0.08	1.00 ± 0.00	0.90 ± 0.00	1.80 ± 0.00
0.20	0.80 ± 0.01	0.93 ± 0.01	0.99 ± 0.00	1.72 ± 0.43	0.79 ± 0.03	0.95 ± 0.01	0.93 ± 0.00	2.67 ± 0.00	0.79 ± 0.08	1.00 ± 0.00	0.90 ± 0.00	1.80 ± 0.00
0.30	0.70 ± 0.01	0.84 ± 0.01	0.95 ± 0.01	1.18 ± 0.02	0.70 ± 0.04	0.92 ± 0.01	0.94 ± 0.00	2.63 ± 0.01	0.71 ± 0.10	1.00 ± 0.00	0.90 ± 0.00	1.80 ± 0.00

human annotations for each image in the CIFAR-10 test set. The human judgments for each image are aggregated into a categorical distribution over the classes, which serves as the oracle label distribution for that image. In all experiments with this dataset, the training data was used to train the models, while the test data was divided into 20% for calibration and 80% for evaluation. As a neural network, we use the PyTorch implementation of ResNet-18, similar in spirit to the setup of Löhr et al. [2025].

ChaosNLI. ChaosNLI [Nie et al., 2020] is an English Natural Language Inference (NLI) dataset in which the task is to classify the textual alignment between a given premise–hypothesis pair into three classes: entailment, contradiction, and neutral. The instances are selected from the development sets of SNLI [Bowman et al., 2015], MNLI [Williams et al., 2018], and AbductiveNLI [Bhagavatula et al., 2019], focusing on examples with high disagreement among labels provided by human annotators. Each premise–hypothesis pair was then annotated by 100 independent humans under strict annotation guidelines, and their responses were aggregated to form the oracle label distribution. We combine the Chaos-SNLI and Chaos-MNLI subsets, resulting in a dataset of 3113 datapoints, of which 80% is used for model training, 10% for calibration, and 10% for evaluation purposes. The premise–hypothesis pairs are embedded in the same manner as in Javanmardi et al. [2024], and a similar fully-connected neural network with 4 hidden layers was used as the base for the ensemble.

QualityMRI. The QualityMRI dataset contains human magnetic resonance images (MRI) with varying levels of diagnostic quality [Obuchowicz et al., 2020]. Each image is evaluated independently by several radiologists, capturing variability in perceived image quality. For our purposes, the task is framed as a binary classification problem, where the aggregated radiologist annotations are interpreted as an oracle label distribution over the two classes [Schmarje et al., 2022]. The dataset is relatively small, comprising 310 instances in total. We use 80% of the data for training, 10% for calibration, and the remaining 10% for evaluation. On this data, we employ the ResNet-18 architecture from the PyTorch torchvision package without pretrained weights as a base for the ensemble.

Appendix G: Extension of Experiments with Partially Valid Credal Sets

In this section, we extend the results of Section 5.2 to the remaining two datasets, namely, ChaosNLI and QualityMRI. Table 4 confirms that what we observed for the case of CIFAR-10 also holds for the other two datasets. Specifically, the conditional coverage satisfaction is always greater than $1-\epsilon$.

Appendix H: Alternative Risk Functions for Calibration

In Section 4.3, when introducing our calibration approach, we specifically considered the risk in (15), namely the indicator that the conditional coverage falls below the target level $1-\alpha$. We chose this risk in order to provide a PAC-style conditional coverage guarantee for our set predictor. However, CRC can be applied with arbitrary risk functions, leading to different types of guarantees. Recall that $\mathcal{D}_{\text{cal}}^{\text{first}} = \{(\boldsymbol{x}_i, \boldsymbol{p}_i)\}_{i=1}^n$ denotes the first-order calibration data, exchangeable with the future test point, and $\mathcal{D}_\dagger = \mathcal{D}_{\text{cal}}^{\text{first}} \cup \{(\boldsymbol{x}_{n+1}, \boldsymbol{p}_{n+1})\}$. Likewise, $\mathcal{D}_{\text{cal}}^{\text{zero}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ denotes the zero-order calibration data, with $\mathcal{D}_+ = \mathcal{D}_{\text{cal}}^{\text{zero}} \cup \{(\boldsymbol{x}_{n+1}, y_{n+1})\}$. When first-order calibration data are available, one may consider the conditional miscoverage risk

$$\mathcal{L}(\boldsymbol{x}_i, \lambda) = 1 - \boldsymbol{b}_i^\lambda \cdot \boldsymbol{p}_i.$$

Alternatively, when only zero-order data are accessible, one may use the marginal miscoverage risk

$$\mathcal{L}(\boldsymbol{x}_i, \lambda) = 1 - \boldsymbol{b}_i^\lambda \cdot \boldsymbol{e}_{y_i},$$

where $\boldsymbol{e}_{y_i}$ denotes the one-hot encoding of y_i. This latter risk can be viewed as a soft extension of classical conformal prediction with miscoverage risk $\mathbb{1}(y_i \notin \mathcal{C}(\boldsymbol{x}_i))$ to the randomized set setting. In Table 5, we provide an overview of the different risk functions, the data they require, and the guarantees they deliver. In Table 7, we compare the performance of calibrated BPS under these risk formulations.

Table 5: Different risk functions, their calibration data requirements, and their guarantees.

Approach Name	Risk $\mathcal{L}(\boldsymbol{x}_i, \lambda)$	Calibration Data	Guarantee
CondSatFirst	$\mathbb{1}\left[\boldsymbol{b}_i^{\lambda} \cdot \boldsymbol{p}_i < 1-\alpha\right]$	first-order data	$\mathrm{P}_{\mathcal{D}_\dagger}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{p}_{n+1} \geq 1-\alpha\right] \geq 1-\beta$
CondSatZero	$\mathbb{1}\left[\boldsymbol{b}_i^{\lambda} \cdot \boldsymbol{e}_{y_i} < 1-\alpha\right]$	zero-order data	$\mathrm{P}_{\mathcal{D}_+}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{e}_{y_{n+1}} \geq 1-\alpha\right] \geq 1-\beta$
MeanCondFirst	$1 - \boldsymbol{b}_i^{\lambda} \cdot \boldsymbol{p}_i$	first-order data	$\mathbb{E}_{\mathcal{D}_\dagger}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{p}_{n+1}\right] \geq 1-\beta$
MargZero	$1 - \boldsymbol{b}_i^{\lambda} \cdot \boldsymbol{e}_{y_i}$	zero-order data	$\mathbb{E}_{\mathcal{D}_+}\left[\boldsymbol{b}_{n+1}^{\lambda^\star} \cdot \boldsymbol{e}_{y_{n+1}}\right] \geq 1-\beta$

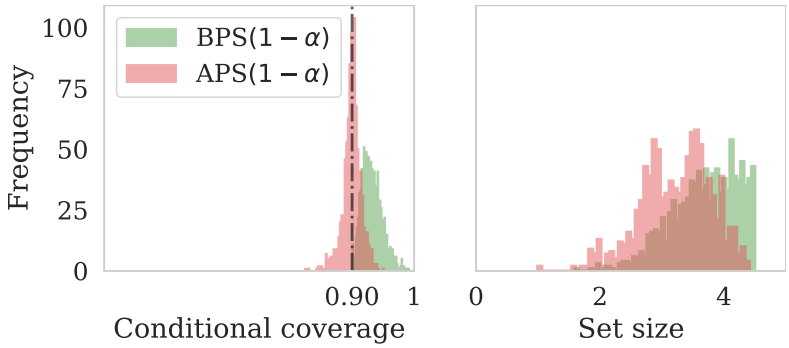


Figure 5: Histograms of the conditional coverage and (expected) set size of BPS$(1-\alpha)$ and APS$(1-\alpha)$ under valid credal sets. Here, $1-\alpha = 0.9$.

Appendix I: Alternative Set Predictors

While in our experiments we used BPS as a natural base set predictor built on top of a credal set predictor, one may ask whether alternative set predictors could be employed. To this end, recall that in the fully valid and partially valid credal set settings (Section 4.1 and Section 4.2), we do not apply any conformal prediction framework. In these cases, BPS alone delivers the desired (exact or PAC-style) conditional coverage guarantee with respect to the true label distribution. Hence, it is not immediately clear what would be gained by replacing BPS with another set predictor.

However, for the case of credal sets with unknown validity (Section 4.3), as we apply CRC to get the PAC-style conditional coverage guarantee, any other set predictor can be used together with the risk defined in (15) and the first-order calibration data to deliver the guarantee. For example, one could aggregate the credal set into a single representative distribution, for instance, by taking the mean over the m distributions defining the credal set, and treat it as a probabilistic classifier. On top of this averaged predictor, methods such as APS could then be applied. While this is a valid construction, it effectively collapses the credal set into a single distribution and therefore no longer explicitly represents epistemic uncertainty. As a consequence, even if such a method satisfies a similar coverage guarantee after calibration, the resulting prediction sets do not adapt to epistemic uncertainty in the same principled way as BPS derived directly from the credal set.

To further investigate this point, we compare the performance of BPS and APS in all three cases. For Cases I and II (valid and partially valid credal sets), we simply replace BPS$(1-\alpha)$ with APS$(1-\alpha)$, i.e., APS with the same nominal coverage level. Figure 5 compares the conditional coverage and set sizes for the case of valid credal sets. It can be seen that, unlike BPS$(1-\alpha)$, APS$(1-\alpha)$ ignores the information contained in the credal sets and violates conditional coverage in a substantial fraction of cases, while producing more efficient (smaller) prediction sets. Table 6 reports the corresponding results for partially valid credal sets. Note that, in this setting, APS is applied to the mean of the ensemble, which is independent of ϵ; hence, unlike BPS, its results do not vary with ϵ, and we report a single row per dataset. A similar pattern emerges: compared to BPS$(1-\alpha)$ as reported in Table 4, APS$(1-\alpha)$ again violates the desired conditional coverage guarantee.

For the case of credal sets with unknown validity, the situation differs. As discussed in Section 4.3, our guarantee can be leveraged for any set predictor, including APS. In particular, calibrated APS can also attain the PAC-style conditional coverage guarantee. Nevertheless, BPS continues to explicitly exploit epistemic uncertainty through the credal set representation, whereas APS operates on a single aggregated distribution and ignores this structure. In Appendix E, we illustrate this difference through toy examples. The full comparison of APS- and BPS-based predictors for this setup under different

Table 6: Performance of APS$(1-\alpha)$ on the partially valid credal sets for different datasets. Since APS operates on the mean predictor, which does not depend on ϵ, the results are identical for all values of ϵ. Here $1-\alpha = 0.9$.

Dataset	Cond. Sat.	Marg. Cvg.	Set Size
CIFAR-10	0.53 ± 0.00	0.90 ± 0.00	1.05 ± 0.00
ChaosNLI	0.56 ± 0.03	0.89 ± 0.01	1.89 ± 0.02
QualityMRI	0.11 ± 0.04	0.52 ± 0.04	1.14 ± 0.03

Table 7: Performance comparison of different set predictors using different calibration approaches on credal sets with unknown validity across three real-world datasets. $1-\alpha = 0.9$ and $1-\beta = 0.9$.

		CIFAR-10			**ChaosNLI**			**QualityMRI**		
base	Calibration	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size
BPS	BPS $(1-\alpha)$	0.52 ± 0.00	0.90 ± 0.00	1.09 ± 0.00	0.61 ± 0.01	0.89 ± 0.01	2.00 ± 0.02	0.33 ± 0.05	0.81 ± 0.02	1.36 ± 0.04
	CondSatFirst	0.90 ± 0.01	0.98 ± 0.00	1.45 ± 0.05	0.91 ± 0.02	0.95 ± 0.01	2.40 ± 0.04	0.94 ± 0.04	0.99 ± 0.01	1.93 ± 0.02
	CondSatZero	0.52 ± 0.00	0.90 ± 0.00	1.09 ± 0.00	0.90 ± 0.01	0.95 ± 0.01	2.35 ± 0.04	0.77 ± 0.10	0.96 ± 0.03	1.81 ± 0.07
	MeanCondFirst	0.75 ± 0.02	0.93 ± 0.00	1.17 ± 0.01	0.70 ± 0.03	0.90 ± 0.01	2.08 ± 0.03	0.83 ± 0.05	0.97 ± 0.01	1.85 ± 0.04
	MargZero	0.34 ± 0.17	0.90 ± 0.00	1.09 ± 0.01	0.71 ± 0.05	0.91 ± 0.02	2.09 ± 0.05	0.66 ± 0.09	0.93 ± 0.03	1.71 ± 0.07
APS	APS $(1-\alpha)$	0.48 ± 0.00	0.89 ± 0.00	0.99 ± 0.00	0.44 ± 0.02	0.84 ± 0.01	1.77 ± 0.01	0.17 ± 0.06	0.76 ± 0.03	1.24 ± 0.04
	CondSatFirst	0.90 ± 0.01	0.99 ± 0.00	1.40 ± 0.07	0.91 ± 0.01	0.95 ± 0.01	2.39 ± 0.03	0.94 ± 0.03	0.99 ± 0.01	1.93 ± 0.02
	CondSatZero	0.48 ± 0.00	0.89 ± 0.00	0.99 ± 0.00	0.89 ± 0.02	0.94 ± 0.01	2.31 ± 0.04	0.73 ± 0.12	0.96 ± 0.03	1.80 ± 0.07
	MeanCondFirst	0.80 ± 0.01	0.93 ± 0.00	1.08 ± 0.01	0.74 ± 0.03	0.90 ± 0.01	2.03 ± 0.03	0.83 ± 0.05	0.97 ± 0.02	1.85 ± 0.03
	MargZero	0.56 ± 0.09	0.90 ± 0.00	1.01 ± 0.01	0.75 ± 0.03	0.91 ± 0.01	2.06 ± 0.05	0.63 ± 0.11	0.93 ± 0.03	1.70 ± 0.09

calibration strategies is given in Table 7.

Remark (APS-based set predictors). We emphasize that although we refer to these methods as APS-based in our experiments, they should not be confused with the original APS approach from the literature [Romano et al., 2020]. The constructions considered here do not appear in prior work; rather, they correspond to our calibration strategies applied to APS as a base set predictor. Only APS$(1-\alpha)$ (APS without calibration) and APS with MargZero calibration (the standard conformal calibration) correspond to the original APS method; the remaining combinations are new.

Remark (comparison of two conformalization strategies). Conformalized credal set predictors [Javanmardi et al., 2024] (as explained in Section 5.2) take any probabilistic classifier together with a set of first-order calibration data and turn it into a credal set predictor that guarantees the true label distribution lies in the credal set with a user-specified probability, hence yielding a partially valid credal set predictor. In Section 4.3, we also showed that, given the risk defined in (15) and first-order calibration data, our method can instead transform a probabilistic classifier directly into a conformal set predictor with a PAC-style conditional coverage guarantee.

A comparison between these two approaches can be seen in our results: Table 4 (BPS$(1-\alpha)$ with $\epsilon = 0.1$) can be compared with Table 7 (APS with CondSatFirst calibration). It can be observed that the second approach, direct conformalization for conditional coverage, yields more efficient (smaller) prediction sets, while the first approach, first constructing conformalized credal sets and then applying optimal BPS with nominal $1-\alpha$, results in larger sets. This is mainly because prediction sets constructed from invalid credal sets may still satisfy conditional coverage, so enforcing credal validity first can be unnecessarily conservative.

Appendix J: Experiments on Out-of-Distribution Data

A domain where epistemic uncertainty arises is when dealing with out-of-distribution (OOD) data for prediction, where data points come from a distribution that deviates from the original training distribution. In such cases, incorporating epistemic uncertainty into set prediction is even more important. To that end, we consider an experimental setup using CIFAR-10-C [Hendrycks and Dietterich, 2019], a corrupted version of CIFAR-10 that is often used to evaluate OOD robustness. Here, as the corruption type, we choose Gaussian noise with severity levels from 1 to 5.

Similar to the in-distribution setting, we apply calibration (if any) solely on in-distribution data. During testing, we use the credal set predictor to provide predictions for OOD input data. Table 8 compares the performance of different set predictors

Table 8: Performance comparison of different set predictors using various calibration approaches on credal sets with unknown validity on CIFAR-10-C (Gaussian noise corruption across different severity levels). $1 - \alpha = 0.9$ and $1 - \beta = 0.9$.

		Severity 1			**Severity 3**			**Severity 5**		
base	Calibration	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size	Cond. Sat.	Marg. Cvg.	Set Size
BPS	BPS $(1-\alpha)$	0.61 ± 0.00	0.87 ± 0.00	1.63 ± 0.01	0.43 ± 0.00	0.60 ± 0.00	2.08 ± 0.01	0.32 ± 0.00	0.48 ± 0.00	2.11 ± 0.01
	CondSatFirst	0.90 ± 0.01	0.96 ± 0.00	2.59 ± 0.13	0.68 ± 0.02	0.77 ± 0.02	3.52 ± 0.21	0.57 ± 0.02	0.67 ± 0.02	3.57 ± 0.20
	CondSatZero	0.61 ± 0.00	0.87 ± 0.00	1.62 ± 0.01	0.43 ± 0.00	0.60 ± 0.00	2.08 ± 0.01	0.32 ± 0.00	0.48 ± 0.00	2.10 ± 0.01
	MeanCondFirst	0.78 ± 0.01	0.90 ± 0.00	1.81 ± 0.02	0.52 ± 0.01	0.65 ± 0.01	2.36 ± 0.03	0.40 ± 0.01	0.53 ± 0.01	2.39 ± 0.03
	MargZero	0.51 ± 0.09	0.87 ± 0.00	1.62 ± 0.01	0.40 ± 0.03	0.60 ± 0.01	2.08 ± 0.02	0.31 ± 0.02	0.48 ± 0.00	2.10 ± 0.02
APS	APS $(1-\alpha)$	0.54 ± 0.00	0.82 ± 0.00	1.26 ± 0.00	0.35 ± 0.00	0.49 ± 0.00	1.48 ± 0.01	0.25 ± 0.00	0.36 ± 0.00	1.45 ± 0.01
	CondSatFirst	0.90 ± 0.01	0.96 ± 0.01	2.46 ± 0.21	0.68 ± 0.03	0.75 ± 0.03	3.32 ± 0.33	0.57 ± 0.04	0.64 ± 0.03	3.30 ± 0.32
	CondSatZero	0.54 ± 0.00	0.82 ± 0.00	1.26 ± 0.00	0.35 ± 0.00	0.49 ± 0.00	1.48 ± 0.01	0.25 ± 0.00	0.36 ± 0.00	1.45 ± 0.01
	MeanCondFirst	0.78 ± 0.01	0.88 ± 0.01	1.48 ± 0.02	0.47 ± 0.01	0.57 ± 0.01	1.81 ± 0.04	0.35 ± 0.01	0.43 ± 0.01	1.79 ± 0.04
	MargZero	0.60 ± 0.06	0.84 ± 0.01	1.31 ± 0.02	0.39 ± 0.02	0.51 ± 0.01	1.56 ± 0.02	0.28 ± 0.01	0.38 ± 0.01	1.53 ± 0.02

under three severity levels: 1, 3, and 5. As the severity increases, the level of corruption also increases, and thus it is expected that the credal sets become larger. This is reflected in the set size of the BPS-based approaches, as they are adaptive to epistemic uncertainty. Furthermore, as expected, BPS outperforms APS in the majority of cases, both with and without calibration, in terms of conditional coverage satisfaction, by incorporating epistemic uncertainty while producing larger sets.