EPISTEMIC LEARNING FROM IMPRECISE ANNOTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Imprecise annotations may support several plausible labelling distributions, yet learning methods often discard this ambiguity into a single predictive distribution. This can obscure what the annotation evidence leaves unresolved. We introduce *epistemic learning from credal supervision*, a framework that uses convex sets of plausible labelling distributions, called credal sets, as supervision and learns sets of compatible predictive distributions. We instantiate the framework with a *pessimistic–optimistic credal classifier* (POCC), which combines a shared backbone with two classification heads trained to minimise worst-case and best-case losses over the available supervisions. Their outputs define a predictive credal set whose spread provides an uncertainty score. We also show how credal labels can practically be obtained through a simple relaxation of existing probabilistic labels, reducing commitment to their precise probability assignments. This construction admits closed-form inner optimisation under cross-entropy loss, enabling efficient training. Assuming the supervision sets contain the true conditional label distributions, and other regularity assumptions, we establish a finite-sample generalisation bound for the averaged predictor with an explicit penalty for supervision imprecision. We evaluate POCC using human annotator disagreement and teacher predictions, alongside label smoothing as a controlled proxy for annotation imprecision. Across these settings, POCC achieves a favourable balance of predictive accuracy, calibration, and uncertainty-based selective classification against competitive baselines.

1 INTRODUCTION

Supervised classifiers usually learn from a single target label or label distribution per instance. Yet annotator disagreement, incomplete label information, and differing interpretations may support several plausible targets (Peterson et al., 2019; Schmarje et al., 2022; Destercke, 2022). Reducing this evidence to a single target can erase distinctions relevant to deciding whether to predict or defer for further assessment. Under majority voting, for example, unanimous agreement and substantial annotator disagreement may produce the same hard label, despite providing different support for that decision.

Existing approaches accommodate annotation imprecision in different ways. Soft labels represent supervision through a single probability distribution, obtained, for example, by aggregating annotations or smoothing hard labels (Geng, 2016; Szegedy et al., 2016). These targets retain information about the relative support for different classes, but they do not explicitly represent which alternative label distributions remain plausible. Another approach represents supervision through a *credal set*: a convex set of admissible label distributions (Lienen & Hüllermeier, 2021a; Lienen et al., 2023). Such sets allow the annotation evidence to leave some probabilities unspecified. However, they are often used to select a training target or define a robust learning objective, after which the model returns a single predictive distribution. The remaining challenge is to carry information about unresolved alternatives from the supervision into the predictions.

Our aim is to develop an *epistemic learner* that represents both what the annotation evidence supports and what it leaves unresolved. We call this objective *epistemic learning from credal supervision*: learning sets of predictive distributions from credal labels. Here, the predictive epistemic uncertainty (EU) of interest concerns ambiguity about which labelling distribution is warranted for a given input, as reflected in the imprecision of the observed annotations. This supervision-induced uncertainty is distinct from outcome variability under a specified distribution and complements un-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

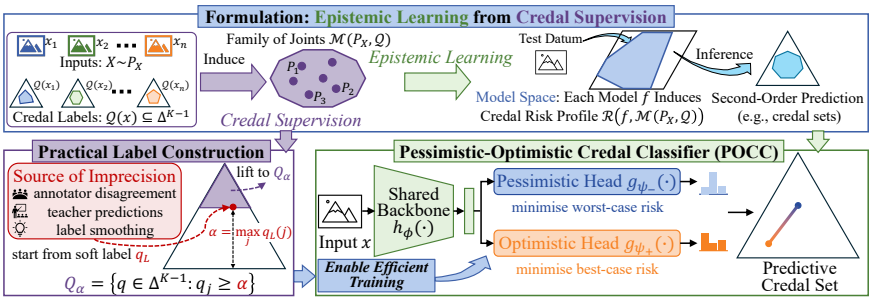


Figure 1: Overview of epistemic learning from credal supervision. Top: the framework relates credal supervision to sets of predictive distributions. Bottom left: practical construction of credal labels from imprecise annotations. Bottom right: POCC uses a shared backbone and pessimistic and optimistic heads to form a predictive credal set.

certainty arising from fitting a model to finite data (Hüllermeier & Waegeman, 2021). Repeated fitting, as in deep ensembles (Lakshminarayanan et al., 2017), captures variation across fitted models but does not, by itself, recover plausible labelling distributions discarded when constructing fixed training targets.

To represent this supervision-induced uncertainty efficiently, we introduce the *pessimistic–optimistic credal classifier* (POCC), whose two prediction heads share a backbone. The pessimistic head minimises the worst-case loss over the distributions admitted by the supervision, while the optimistic head minimises the best-case loss. Their outputs and all mixtures between them form a predictive credal set. The separation between the heads supplies an uncertainty score, while their average provides a probability prediction. Figure 1 provides an overview of our framework and proposed method. We summarise our contributions below:

- **A framework for epistemic learning from credal supervision.** We formalise learning from sets of admissible label distributions and identify the representation of unresolved annotation evidence at prediction time as a central learning objective.
- **An efficient classifier with a learning-theoretic analysis.** We derive closed-form solutions for POCC's inner optimisation problems under cross-entropy loss and the proposed credal-label construction. Assuming the supervision sets contain the true conditional label distributions, and under appropriate boundedness conditions, we establish a finite-sample generalisation bound for the averaged predictor with an explicit penalty for supervision imprecision.
- **An empirical evaluation across three supervision settings.** We construct credal labels from human annotator disagreement and teacher predictions, and use label smoothing as a controlled proxy for annotation imprecision. Across the evaluated settings, POCC achieves a favourable overall balance among predictive accuracy, calibration, and uncertainty-based selective classification compared with competitive baselines.

2 Background

Classical Supervised Classification. Consider a K-class classification problem with input space $\mathcal{X} \subseteq \mathbb{R}^d$, label space $\mathcal{Y} = \{1, \ldots, K\}$, and the $(K-1)$-dimensional probability simplex $\Delta^{K-1} = \{q \in \mathbb{R}^K_{\geq 0} : \sum_{k=1}^K q_k = 1\}$. Assuming data (X, Y) are drawn i.i.d. from a true but unknown joint distribution P_0 over $\mathcal{X} \times \mathcal{Y}$, the standard objective is to learn a first-order probabilistic classifier $f : \mathcal{X} \to \Delta^{K-1}$ minimising the population risk

$$R_0(f; P_0) = \mathbb{E}_{(X,Y)\sim P_0}\big[\ell(f(X), Y)\big], \tag{1}$$

where $\ell : \Delta^{K-1} \times \mathcal{Y} \to \mathbb{R}$ denotes the loss function (e.g., cross-entropy). In practice, given a finite dataset $\{(x_i, y_i)\}_{i=1}^N$, the empirical risk $\hat{R}(f; \hat{P}_N) = \frac{1}{N}\sum_{i=1}^N \ell(f(x_i), y_i)$ is minimised instead. To assess performance on unseen data, learning theory studies the gap between empirical and population risks, known as the *generalisation gap*. Under suitable assumptions on the loss and model complexity, bounds on this gap establish when increasing data and improving optimisation bring the learned predictor's risk towards the best attainable in its model class (Shalev-Shwartz & Ben-David, 2014; Bartlett & Mendelson, 2002), motivating our analysis in Section 4. For a

distributional target $q \in \Delta^{K-1}$, let $\ell(p, q) = \sum_k q_k \, \ell(p, k)$, which recovers $\ell(p, k)$ at the one-hot target $q = e_k$; For cross-entropy, this gives $\ell(p, q) = -\sum_{k=1}^{K} q_k \log p_k$.

Learning from Imprecise Annotations. We consider single-label classification in which the annotation evidence may leave several conditional labelling distributions plausible. This differs from label-noise formulations, which model discrepancies between observed and underlying clean labels (Song et al., 2023; 2025), and from multi-label learning, where several classes may apply simultaneously (Tarekegn et al., 2024). A missing annotation supplies no direct label constraint, although missing-label recovery lies outside our scope.

Soft labels and label distribution learning summarise class support in a single probability vector (Geng, 2016; Schmarje et al., 2022), while label smoothing deliberately softens hard targets (Szegedy et al., 2016). Credal labels retain a set of admissible distributions, allowing the supervision to leave some target probabilities unresolved (Lienen & Hüllermeier, 2021b). They have also been used as pseudo-labels in semi-supervised and self-supervised learning (Lienen et al., 2023; Rodriguez et al., 2023; Bordini et al., 2024).

Theoretical work has studied learning guarantees under different credalised settings (Campagner, 2023; Caprio et al., 2024), such as learning from multiple data generating sources. Closest to our work is Venkatesh et al. (2026), where they developed a structured credal framework that distinguishes uncertainty in label information from uncertainty in the input distribution. Much of this work learns ordinary probabilistic predictors; our interest also concerns how supervision imprecision informs predictive epistemic uncertainty.

Representing Predictive Epistemic Uncertainty. A probability vector describes predictions over class outcomes. Uncertainty about which probability vector is warranted can itself be represented by a distribution over probability vectors or by a set of plausible vectors (Hüllermeier & Waegeman, 2021; Wang et al., 2026b). We use *second-order representation* to refer broadly to these richer descriptions of predictive uncertainty. A credal set, in particular, specifies which distributions are admissible without assigning probabilities to its members.

Bayesian and ensemble methods obtain such representations through uncertainty over model parameters or variation across fitted models (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Hobbhahn et al., 2022). Evidential and related approaches learn representations of uncertainty over class probabilities directly (Sensoy et al., 2018; Ni et al., 2026), while credal predictors return sets of plausible probability vectors (Wang et al., 2024; Hofman et al., 2026). These predictive representations do not, by themselves, specify how the constraints supplied by imprecise annotations should inform predictive epistemic uncertainty.

3 Epistemic Learning from Credal Supervision

Imprecise annotations constrain plausible conditional labelling distributions. To carry these constraints into prediction, we first show that they induce a credal set of joint distributions and a range of population risks. The endpoints then define complementary learning objectives.

For each input $x \in \mathcal{X}$, let $Q(x) \subseteq \Delta^{K-1}$ contain the conditional labelling distributions compatible with the annotation evidence. A conventional Markov kernel assigns one conditional distribution to each input; we extend this representation by allowing a set of distributions.

Definition 3.1 (Credal Markov Kernel)**.** A *credal Markov kernel* is a measurable set-valued map $Q : \mathcal{X} \rightrightarrows \Delta^{K-1}$ such that $Q(x)$ is non-empty, closed, and convex for every $x \in \mathcal{X}$.

Hard and soft labels correspond to singleton sets: $Q(x) = \{e_k\}$ for class k, where e_k places all mass on k, and $Q(x) = \{q\}$ for a soft target q. Non-singleton sets leave several conditional laws admissible. Fixing the input marginal P_X induces the following joint credal set and range of risks.

Proposition 3.2 (Induced Joint Credal Set and Risk Profile)**.** *Assume that $\mathcal{X}$ is a Polish space—a space admitting a complete separable metric—equipped with its Borel σ-algebra. For a fixed input marginal P_X and a credal Markov kernel Q, the family*

$$\mathcal{M}(P_X, Q) = \big\{ P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : P(\,\cdot \times \mathcal{Y}) = P_X,\ P_{Y|X=x} \in Q(x) \text{ for } P_X\text{-a.e. } x \big\}$$

is non-empty, convex, and weakly compact, and hence is a closed convex credal set. Here, $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the probability distributions on the input–label space. Suppose a measurable predictor f

has measurable per-class losses. If $\mathbb{E}_X[\max_k |\ell(f(X), k)|] < \infty$ *for* $X \sim P_X$*, its* credal risk profile *is the compact interval*

$$\mathcal{R}(f, \mathcal{M}) := \{\mathbb{E}_P[\ell(f(X), Y)] : P \in \mathcal{M}\} = [R_+(f), R_-(f)], \qquad \mathcal{M} := \mathcal{M}(P_X, Q), \tag{2}$$

where

$$\begin{aligned} R_+(f) &= \min_{P \in \mathcal{M}(P_X, Q)} E_P\big[\ell(f(X), Y)\big] = \mathbb{E}_X\big[\min_{q \in Q(X)} \ell(f(X), q)\big], \\ R_-(f) &= \max_{P \in \mathcal{M}(P_X, Q)} E_P\big[\ell(f(X), Y)\big] = \mathbb{E}_X\big[\max_{q \in Q(X)} \ell(f(X), q)\big]. \end{aligned} \tag{3}$$

Both endpoints are attained by distributions in $\mathcal{M}(P_X, Q)$*.*

The joint credal set supports a behavioural interpretation: its members are permissible probability models for evaluating decisions through expected loss, without commitment to a single model (Levi, 1980). This formulation also connects to the ambiguity sets used in distributionally robust optimisation (DRO) (Ben-Tal et al., 2013; Duchi et al., 2021; Chen et al., 2026). Whereas standard DRO formulations minimise worst-case expected loss to seek robustness, our credal formulation retains the full range of expected losses induced by the permissible models, without privileging either endpoint as the sole learning criterion. In our setting, the input marginal remains fixed, and the geometry of the joint credal set is determined by the conditional annotation constraints $Q(x)$. The proof is provided in Appendix B.1.

3.1 Pessimistic–Optimistic Credal Classifier

One natural approach to learning from credal supervision is to train a separate predictor for each admissible conditional labelling law. This leads to an infinite-task learning problem (Brault et al., 2019; Chen et al., 2024; Singh et al., 2024): the family of laws is generally uncountable, and approximating it through sampled tasks can introduce substantial computational overhead. We instead introduce the *pessimistic–optimistic credal classifier* (POCC), which uses the two endpoints of the credal risk profile to define complementary learning objectives.

Pessimistic and Optimistic Learning. The upper risk R_- and lower risk R_+ measure worst- and best-case performance over the set of admissible joint distributions, yielding classical pessimistic and optimistic learning criteria (Guillaume et al., 2017; Hüllermeier et al., 2019). POCC minimises these risks with two predictors and retains both outputs in a predictive credal set. This represents the consequences of unresolved annotation evidence, rather than selecting a single compromise between robustness and optimism.

Architecture and Training. As illustrated in Figure 1, POCC combines a shared backbone with two classification heads. Let h_ϕ denote the backbone and g_{ψ_-} and g_{ψ_+} the pessimistic and optimistic heads. For $\theta = (\phi, \psi_-, \psi_+)$, their predictive distributions are

$$p_{-,\theta}(x) := f_{-,\theta}(x) = g_{\psi_-}(h_\phi(x)), \qquad p_{+,\theta}(x) := f_{+,\theta}(x) = g_{\psi_+}(h_\phi(x)), \tag{4}$$

with both outputs taking values in Δ^{K-1}. Given training data $\{(x_i, Q_i)\}_{i=1}^N$, where $Q_i := Q(x_i)$, we jointly train the two heads by solving

$$\min_\theta \ \frac{1}{N} \sum\nolimits_{i=1}^N \Big[\max_{q \in Q_i} \ell(p_{-,\theta}(x_i), q) + \min_{q \in Q_i} \ell(p_{+,\theta}(x_i), q)\Big]. \tag{5}$$

For each instance, the inner maximisation selects the least favourable admissible label distribution for the pessimistic head, while the inner minimisation selects the most favourable one for the optimistic head. Both heads minimise their respective losses, and both contribute to learning the shared representation.

This objective applies whenever the inner extrema over the credal labels can be evaluated. For the construction introduced in Section 3.2, both admit closed-form solutions under cross-entropy loss, allowing efficient training without iterative inner optimisation.

Prediction and Epistemic Uncertainty. To retain both learned views at prediction time, we return the set of all mixtures of the two head outputs:

$$C_\theta(x) = \{(1 - \lambda) p_{-,\theta}(x) + \lambda p_{+,\theta}(x) : \lambda \in [0, 1]\}. \tag{6}$$

This line segment is the smallest closed convex set that contains both predictions, called their convex hull, and it collapses to a singleton when they agree (Huber, 1992). When a single predictive distribution is required, we use the midpoint and its most probable class:

$$\bar{f}_\theta(x) = \big(p_{-,\theta}(x) + p_{+,\theta}(x)\big)/2, \qquad \widehat{y}_\theta(x) = \arg\max_k \bar{f}_{\theta,k}(x).$$

We use the spread of $C_\theta(x)$ as an epistemic uncertainty score reflecting disagreement between the two learning criteria. Specifically, we adopt the *Maximum Mean Imprecision* (MMI) under the total variation distance (Chau et al., 2025a; Gonzalez-Garcia et al., 2026). For this line segment, MMI reduces to the total variation distance between its endpoints:

$$\mathrm{MMI}(x) = \frac{1}{2}\sum_{k=1}^{K} \left|p_{+,\theta,k}(x) - p_{-,\theta,k}(x)\right|. \tag{7}$$

The score is available directly from the two head outputs and requires no further optimisation. Appendix I.1 provides the derivation, while Appendix I compares it with classical entropy-based alternatives (Abellán et al., 2006).

3.2 Practical Credal-Label Construction and Efficient Optimisation

Credal supervision allows annotation evidence to leave some probabilities unspecified. For example, a statement such as "at least 70% probability of cat" naturally defines a set of compatible labelling distributions. Common supervision signals instead often provide a single reference distribution $q_L(x) \in \Delta^{K-1}$, obtained from annotator frequencies, teacher predictions, or smoothed labels. We show how to obtain credal labels through a simple relaxation of these readily available targets, reducing commitment to their precise probability assignments while retaining information useful for prediction.

Following Lienen & Hüllermeier (2021a), we retain the probability of the most supported class as a lower bound and leave the allocation of the remaining mass unresolved. Let $j(x) \in \arg\max_k q_L(k \mid x)$, with ties resolved by a fixed rule, and write $\alpha(x) = q_L(j(x) \mid x)$. The resulting credal label is

$$\mathcal{Q}_\alpha(x) := \left\{q \in \Delta^{K-1} : q_{j(x)} \geq \alpha(x)\right\}. \tag{8}$$

By construction, $\mathcal{Q}_\alpha(x)$ contains $q_L(x)$, as illustrated in Figure 1. The relaxation therefore preserves the original target as an admissible distribution while allowing alternatives to its precise probability assignments.

Interpreting Supervision Imprecision. To understand what this relaxation represents, observe that

$$\mathcal{Q}_\alpha(x) = \alpha(x)e_{j(x)} + \big(1-\alpha(x)\big)\Delta^{K-1}, \tag{9}$$

where $e_{j(x)}$ places all probability mass on class $j(x)$. Every admissible distribution reserves mass $\alpha(x)$ for this class and leaves the remaining mass unrestricted, including allocation back to the same class. This is an ϵ-contamination set with $\epsilon(x) = 1 - \alpha(x)$: the mass originally assigned outside the most supported class determines how much probability mass is left free to be reallocated.

The remaining imprecision is quantified by the largest total variation distance between members of the set: for $K \geq 2$, $\mathrm{diam}_{\mathrm{TV}}\big(\mathcal{Q}_\alpha(x)\big) := \sup_{q,r\in\mathcal{Q}_\alpha(x)} \mathrm{TV}(q,r) = 1 - \alpha(x)$. For fixed $j(x)$, decreasing $\alpha(x)$ enlarges the set. A one-hot target $q_L(x) = e_{j(x)}$ gives $\alpha(x) = 1$ and the singleton $\{e_{j(x)}\}$. Although $q_L(x)$ belongs to the set by design, coverage of the true conditional distribution remains an assumption of Section 4.

Closed-Form Inner Optimisation. The same mixture representation makes the POCC training objective efficient to evaluate. Cross-entropy is linear in the target distribution, so each head's extremal target is obtained by assigning the unrestricted mass to a single class.

Proposition 3.3 (Closed-Form Extreme Label Distributions)**.** *Fix an input x, abbreviate $j = j(x)$ and $\alpha = \alpha(x)$, and let p_- and p_+ denote the two head predictions with strictly positive coordinates. Choose $k_{\min} \in \arg\min_k p_{-,k}$ and $k_{\max} \in \arg\max_k p_{+,k}$. Under cross-entropy loss, an inner maximiser for the pessimistic head and an inner minimiser for the optimistic head are*

$$q_-^* = \alpha e_j + (1-\alpha)e_{k_{\min}}, \qquad q_+^* = \alpha e_j + (1-\alpha)e_{k_{\max}}. \tag{10}$$

These expressions also cover $k_{\min} = j$ or $k_{\max} = j$, in which case the corresponding target is e_j.

The pessimistic target places the unrestricted mass on the class its head considers least probable, thereby maximising the loss. The optimistic target places this mass on the class its head considers most probable, thereby minimising the loss. A proof is provided in Appendix B.2. Each inner problem requires an $\mathcal{O}(K)$ scan, with no sampling or iterative optimisation. The dependence on the number of classes matches ordinary cross-entropy training; the additional model cost comes primarily from the second head.

4 Learning Guarantees under Credal Supervision

Our analysis examines when imprecise supervision supports informative probabilistic prediction. As a first step, we consider *valid but imprecise supervision*: the annotation constraints may leave the true conditional labelling distribution unresolved, but do not exclude it. Under suitable regularity conditions, we establish a finite-sample cross-entropy risk bound for POCC's averaged predictor, making explicit the contribution of supervision imprecision.

Coverage of the True Conditional Distribution. Let $p_0(x) = P_0(Y = \cdot \mid X = x)$ denote the true conditional labelling distribution. We formalise supervision validity through the following assumption.

Assumption 4.1 (Almost-sure Conditional Coverage)**.** The credal supervision satisfies $p_0(x) \in Q(x)$ for P_X-almost every x.

Coverage is a *well-specification assumption on the supervision*, allowing us to isolate the effect of annotation imprecision from that of misspecified constraints. It relaxes exact distributional supervision, $Q(x) = \{p_0(x)\}$, to set membership. This is distinct from the usual sampled-label assumption: the latter specifies how observed labels are drawn from the target conditional distribution, whereas coverage specifies the validity of the supplied annotation constraints. In particular, coverage neither requires deterministic outcomes nor follows merely from i.i.d. sampling. When labels are deterministic, $p_0(x) = e_{y^*(x)}$, a correct hard annotation is recovered as the singleton $Q(x) = \{e_{y^*(x)}\}$. More general supervision models that account for stochastic annotation mechanisms and coverage violations, for example through random-set formulations, are left to future work.

By ensuring $P_0 \in \mathcal{M}(P_X, Q)$, coverage connects the credal objectives to the true risk $R_0(f) := \mathbb{E}_X[\ell(f(X), p_0(X))]$, yielding $R_+(f) \le R_0(f) \le R_-(f)$. Coverage alone, however, does not ensure informative supervision: taking $Q(x) = \Delta^{K-1}$ always satisfies the assumption but leaves the conditional distribution unrestricted. We therefore quantify supervision imprecision through the total-variation diameter $d_Q(x) := \sup_{q,r \in Q(x)} \frac{1}{2}\|q - r\|_1$.

Generalisation of the Averaged Predictor. To relate the training objective in eq. (5) to this risk, we divide it by two, leaving its minimisers unchanged. Define

$$j_\theta(x) := \big(\max_{q \in Q(x)} \ell(f_{-,\theta}(x), q) + \min_{q \in Q(x)} \ell(f_{+,\theta}(x), q)\big)/2, \tag{11}$$

and let $\widehat{J}_{Q,n}(\theta) := n^{-1} \sum_{i=1}^n j_\theta(X_i)$. Suppose that training returns a parameter $\widehat{\theta}_n$ satisfying $\widehat{J}_{Q,n}(\widehat{\theta}_n) \le \inf_{\theta \in \mathbf{\Theta}} \widehat{J}_{Q,n}(\theta) + \eta_n$, where $\eta_n \ge 0$ measures the optimisation error. Its averaged prediction is $\widehat{\bar{f}}_n = (f_{-,\widehat{\theta}_n} + f_{+,\widehat{\theta}_n})/2$. To assess this predictor against an ordinary probabilistic model of comparable capacity, we use the predictors that the architecture can represent with identical heads:

$$\mathcal{F}_\Delta := \left\{ f : \exists \theta_f \in \mathbf{\Theta} \text{ such that } f_{-,\theta_f} = f_{+,\theta_f} = f \right\}.$$

The learned heads may differ; only the comparison class is restricted in this way.

Theorem 4.2 (Generalisation under Credal Coverage)**.** *Assume almost-sure conditional coverage, a fixed credal kernel Q, and inputs $X_1, \dots, X_n$ drawn independently from P_X. Suppose that $\mathcal{F}_\Delta$ is non-empty and, on one common measurable set of inputs with P_X-probability one, cross-entropy is bounded by B for all parameters, both heads, and all admissible targets, while $\max_k[-\log f_k(x)] - \min_k[-\log f_k(x)] \le L$ for every head prediction. Then, with probability at least $1 - \delta$,*

$$R_0(\widehat{\bar{f}}_n) - \inf_{f \in \mathcal{F}_\Delta} R_0(f) \le 4\mathfrak{R}_n(\mathcal{J}_Q) + 2B\sqrt{\frac{\log(2/\delta)}{2n}} + \eta_n + L\,\mathbb{E}[d_Q(X)], \tag{12}$$

where $\mathcal{J}_Q := \{j_\theta : \theta \in \mathbf{\Theta}\}$, and $\mathfrak{R}_n(\mathcal{J}_Q)$ measures the capacity of this loss class through its expected absolute Rademacher complexity, defined in Appendix B.3.

The bound separates the effects of model complexity, finite-sample variation, imperfect optimisation, and annotation imprecision. The final term is specific to the unresolved supervision: even when coverage holds, different admissible conditional distributions can assign different risks to the same predictor. Its contribution decreases as the credal labels become more precise and vanishes when $Q(x) = \{p_0(x)\}$ almost surely. The proof uses convexity of cross-entropy to compare the averaged predictor with its heads, coverage and diameter bounds to control discrepancies between true and credal risks, and uniform convergence to relate empirical and population objectives; see Appendix B.3.

Specialisation to our Credal Construction. For $Q_{\alpha,j}(x) = \{q \in \Delta^{K-1} : q_{j(x)} \geq \alpha(x)\}$ from eq. (8), coverage has the explicit interpretation $p_{0,j(x)}(x) \geq \alpha(x)$ almost surely: the probability lower bound supplied by the annotation must be valid for the true conditional distribution. If all logits in the model class are bounded in absolute value by A, substituting $B = \log K + 2A$, $L = 2A$, and $d_Q(x) = 1 - \alpha(x)$ into eq. (12) gives the imprecision penalty $2A\,\mathbb{E}[1 - \alpha(X)]$. Appendix B.3.4 provides the complete specialised bound and its derivation.

This specialisation makes the relationship between validity and precision explicit. Increasing $\alpha(x)$ narrows the credal label and reduces the imprecision penalty, but also strengthens the coverage requirement. Containing the reference target $q_L(x)$ does not by itself establish coverage of $p_0(x)$. The guarantee therefore applies when the annotation constraints are valid, and quantifies the effect of the uncertainty they leave unresolved. It concerns the cross-entropy risk of the averaged predictor; coverage or calibration of the predictive credal set requires a separate analysis.

5 Experiments

Evaluation and Baselines. We study annotation imprecision from two sources—annotators' disagreement and teacher predictions in a knowledge-distillation setting—and include label smoothing as a controlled proxy. We report test accuracy (ACC), expected calibration error (ECE) (Guo et al., 2017), and the usefulness of EU estimates for *selective classification*, in which predictions with high uncertainty are rejected. The area under the accuracy–rejection curve (AUARC) (Jaeger et al., 2023) summarises accuracy on retained predictions across rejection rates. Higher ACC and AUARC and lower ECE are better. For EU evaluation, we adopt selective classification rather than out-of-distribution (OOD) detection as the latter targets distributional shift in input space, whereas our EU stems from label-space annotation imprecision. OOD evaluation was also recently criticised as an inappropriate evaluation for EU estimation (Li et al., 2025) (see Appendix J). To summarise performance, the *balanced quality score* (BQS) averages ACC, $1-$ECE, and AUARC after separate min-max normalisation to $[0, 1]$ across methods for each seed. All experiments are repeated over 10 independent seeds (seed$=1,\ldots,10$).

Table 1: CIFAR-10H results (%). Best in bold; second best underlined.

Method	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
EDL	85.0±0.4	9.5±0.3	96.7±0.1	33.1±8.3
LbBnn	**87.0±0.8**	6.6±0.3	95.9±0.3	54.8±6.3
Decali	84.3±1.0	1.0±0.2	94.9±0.5	35.4±4.5
DAPPr	86.6±0.3	6.9±0.2	**97.2±0.1**	68.3±5.9
Ours	**87.0±0.4**	2.9±0.3	97.1±0.2	**86.0±6.1**
MCDO	85.6±0.7	**0.9±0.1**	96.6±0.3	71.0±10.6
DE	**87.0±0.3**	4.3±0.2	96.9±0.1	77.6±5.4

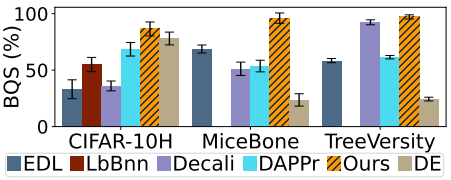


Figure 2: BQS comparison under annotator disagreement on distinct datasets.

Our work introduces a new problem setting, *epistemic learning from credal supervision*, with POCC as one instantiation. To our knowledge, no existing method is designed specifically for this setting. To provide informative comparisons, we adapt established methods for *epistemic uncertainty quantification* to the same supervision sources. Specifically, the baselines are trained on the reference distributional target q_L, while POCC uses the credal label constructed from that same target via eq. (8). This design gives all methods access to the same underlying annotation information while accommodating their different supervision representations. Specifically, we consider four single-pass second-order predictors: evidential deep learning (EDL) (Sensoy et al., 2018), Laplace bridge with Bayesian deep networks (LbBnn) (Hobbhahn et al., 2022), efficient credal predictions via decalibration (Decali) (Hofman et al., 2026), and Dirichlet-approximated possibilistic posterior predictions (DAPPr) (Ni et al., 2026). We also include inference-heavy but widely used baselines:

Monte Carlo dropout (MCDO) (Gal & Ghahramani, 2016) with five forward passes and deep ensembles (DE) (Lakshminarayanan et al., 2017) with five members. To ensure a fair comparison, batch size, training epochs, optimiser, and learning rate schedule are matched across methods. Implementation details are in Appendix C.

Annotator Disagreement. For class labels $y_1, \ldots, y_S$ supplied by S annotators, we use their empirical distribution $q_L = S^{-1} \sum_{s=1}^{S} e_{y_s}$ as the distributional training target q_L.

Setup. CIFAR-10H (Peterson et al., 2019) supplies about 50 annotations per image for the CIFAR-10 test set. We train all methods from scratch with a ResNet-18 (He et al., 2016) backbone, and evaluate them on the last 10,000 samples of the original CIFAR-10 training set to avoid data leakage. On MiceBone and TreeVersity (Schmarje et al., 2022), we use DenseNet-121 (Huang et al., 2017) pre-trained on ImageNet. For these two datasets, MCDO and LbBnn are excluded due to the lack of a compatible pre-trained model and out-of-memory errors on an NVIDIA RTX 5090, respectively.

Results. Table 1 reports performance on CIFAR-10H, with BQS comparison across datasets in Figure 2 and additional results in Table A4. POCC achieves the highest BQS across all three datasets. Although MCDO and Decali obtain lower ECE, this comes at the cost of substantially worse ACC and AUARC, reflected in their markedly lower BQS.

Teacher Predictions. In the knowledge distillation setting (Hinton et al., 2015), a pre-trained teacher's soft predictions serve as distributional labels q_L.

Setup. We use publicly available teachers (Chen, 2025) pre-trained on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). To assess the generalisability of our findings, we consider three teacher architectures: ResNet-20, ShuffleNetV2-0.5× (Ma et al., 2018), and MobileNetV2-x0-5 (Sandler et al., 2018). Student models adopt ResNet-18 for CIFAR-10 and WRN-28-4 (Zagoruyko & Komodakis, 2016) for CIFAR-100, with temperature $T = 2.5$ (Hinton et al., 2015). We additionally ablate on unscaled teacher predictions ($T = 1$). Implementation details are provided in Appendix C.3.2.

Results. Table 2 and Figure 3 show that POCC achieves the best mean ACC, ECE, and BQS across all three teacher architectures and both datasets under temperature-scaled predictions ($T = 2.5$). Under unscaled predictions ($T = 1$), POCC achieves the best mean BQS across all three teacher architectures and both datasets. Detailed results for all settings are provided in Tables A5 and A7.

Table 2: Teacher-prediction results (%) with ResNet-20. Best in bold; second best underlined.

Method	CIFAR-10				CIFAR-100			
	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
EDL	93.4±0.2	9.0±0.2	99.1±0.1	51.6±3.8	33.1±5.7	23.3±3.9	65.8±6.1	30.2±2.9
LbBnn	95.1±0.1	28.1±0.2	99.2±0.0	54.2±3.3	73.1±0.2	63.1±0.2	63.4±0.4	31.9±2.4
Decali	93.5±0.1	8.5±0.1	98.0±0.1	29.7±2.5	71.0±0.2	27.8±0.2	73.3±0.3	64.2±1.8
DAPPr	93.4±0.2	14.6±0.2	99.0±0.0	42.8±3.1	72.5±0.4	52.8±0.3	78.3±0.3	55.1±1.6
Ours	**95.5**±0.1	**3.0**±0.1	**99.5**±0.0	**99.9**±0.3	**77.2**±0.2	**10.6**±0.4	**90.6**±0.2	**100.0**±0.0
MCDO	93.5±0.2	8.5±0.2	99.0±0.1	50.2±4.8	71.1±0.3	28.2±0.3	83.2±0.7	75.5±1.3
DE	94.2±0.1	9.5±0.1	99.2±0.0	65.3±3.0	72.7±0.3	29.9±0.3	86.5±0.3	79.5±0.7

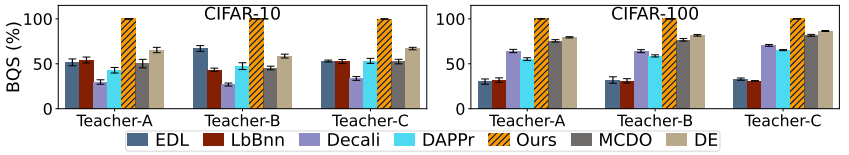


Figure 3: BQS across teachers: (A) ResNet-20; (B) ShuffleNetV2-0.5×; (C) MobileNetV2-x0-5.

Label Smoothing. Label smoothing (Szegedy et al., 2016) redistributes a fraction ϵ of a one-hot target's probability mass uniformly across the other classes. Unlike annotator disagreement or teacher predictions, which reflect instance-dependent ambiguity, it applies a fixed perturbation across all samples. It thus serves as a *proxy* for annotation imprecision rather than a direct instantiation of it.

Setup. We use ResNet-18 on CIFAR-10 and WRN-28-4 on CIFAR-100, with label smoothing parameter $\epsilon \in \{0.05, 0.1, 0.15, 0.2\}$. Configurations are detailed in Appendix C.3.3.

Results. Table 3 reports results for $\epsilon = 0.05$, with BQS across all ϵ visualized in Figure 4 and detailed results provided in Table A6. POCC achieves the highest mean BQS in seven of eight settings; the sole exception is $\epsilon = 0.1$ on CIFAR-10, where it ranks second behind DE (5 members), with the gap attributable to marginal differences in accuracy and AUARC. Statistical significance tests (Table A10, Appendix E) confirm that POCC ranks first statistically across all settings.

Table 3: Label-smoothing results (%) at $\epsilon = 0.05$ over 10 runs. Best in bold; second best underlined.

Method	CIFAR-10				CIFAR-100			
	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
EDL	93.4±0.3	9.1±0.2	99.1±0.1	48.6±4.3	26.3±1.2	18.7±1.0	58.5±1.6	24.3±0.6
LbBnn	95.4±0.2	29.9±0.3	99.1±0.1	53.6±2.6	79.3±0.3	61.6±0.5	74.5±0.9	47.7±1.1
Decali	93.9±0.3	3.6±0.1	98.3±0.2	39.0±5.2	76.2±0.3	3.5±0.4	90.7±0.3	94.0±0.7
DAPPr	93.9±0.3	8.9±0.2	99.1±0.1	58.5±6.4	75.1±0.4	35.3±0.4	91.9±0.2	76.5±0.4
Ours	**95.5±0.1**	**1.5±0.1**	99.2±0.1	**90.6±4.7**	79.3±0.2	**2.7±0.2**	93.4±0.1	**99.0±0.3**
MCDO	93.7±0.2	3.5±0.1	99.1±0.1	60.2±6.3	76.1±0.3	3.8±0.4	91.8±0.2	94.9±0.6
DE	95.2±0.1	5.5±0.1	**99.4±0.0**	89.7±2.3	**80.3±0.2**	9.1±0.2	**93.8±0.1**	96.4±0.2

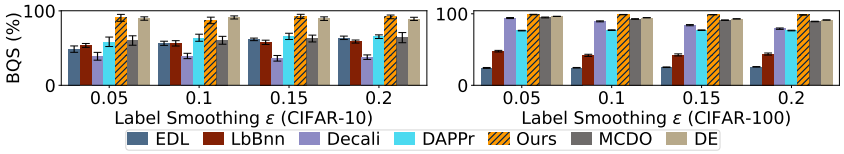


Figure 4: BQS comparison across label-smoothing strengths ϵ.

Discussion. Across the three supervision settings, POCC achieves a consistently favourable balance among predictive accuracy, calibration, and uncertainty ranking, obtaining the highest mean BQS in most evaluated configurations. Individual baselines remain competitive and sometimes perform better on particular metrics, but their relative strengths vary across settings.

These findings are consistent with the motivation for explicitly modelling annotation imprecision. Although a distributional target q_L retains information about relative class support, treating it as fixed does not explicitly represent uncertainty about which alternative labelling distributions remain plausible. POCC incorporates these alternatives into learning through complementary pessimistic and optimistic objectives, and retains both heads' predictions in a predictive credal set. This provides a mechanism through which unresolved supervision can influence the predictions and their associated uncertainty score. The observed balance across metrics suggests that this treatment of supervision ambiguity is practically useful, even when POCC does not achieve the best performance on every individual metric.

Additional Analyses. We provide further empirical evidence that supports the merit of our framework and method. Appendix E reports pairwise one-sided Wilcoxon signed-rank tests at the 5% significance level, confirming that POCC is statistically ranked first in balanced performance across distinct settings. Appendix F compares POCC with baselines trained on collapsed hard labels, also showing the best trade-off among ACC, calibration, and EU quantification. Appendices G and H assess balanced performance via normalised AUARC and the EU–Prediction Error correlation, respectively, where POCC consistently achieves the best balance score. Appendix I ablates POCC across different EU measures, demonstrating robustness to the choice of measure.

6 Conclusion and Future Work

We introduced *epistemic learning from credal supervision*, which uses sets of admissible labelling distributions to retain unresolved annotation information in prediction. POCC implements this framework with complementary pessimistic and optimistic heads and closed-form inner optimisation. Under conditional coverage and boundedness assumptions, its averaged predictor admits a finite-sample risk bound with an explicit imprecision term. Across annotator disagreement, teacher predictions, and label smoothing as a controlled proxy, POCC achieves a favourable overall balance of accuracy, calibration, and selective classification.

Limitations and Future Work. POCC is one instantiation of this framework. Future work will explore alternative predictors and credal-label constructions, and extensions to large foundation models. Our analysis provides a first guarantee under almost-sure conditional coverage, separating the effect of supervision imprecision from that of invalid annotation constraints. This condition need not hold for credal labels constructed from finite annotations or imperfect teacher predictions. A natural extension is to model credal labels as random sets generated by a stochastic annotation mechanism, replacing almost-sure coverage with probabilistic coverage or compatibility in conditional expectation. Such an analysis would account for annotation variability and controlled misspecification alongside the imprecision retained by the supervision.

AI USE STATEMENT

In this work, we used generative AI tools to help polish the writing and check mathematical proofs. We have reviewed all AI-assisted work.

REPRODUCIBILITY STATEMENT

Experimental configurations, hyperparameters, and dataset preprocessing procedures are detailed in Appendix C. The datasets used in this work are drawn from existing publicly available benchmarks. To further support reproducibility, the core implementation codes are included in the supplementary materials and will be released publicly upon acceptance under a licence permitting free use for research purposes.

REFERENCES

Joaquín Abellán, George J Klir, and Serafín Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1):29–44, 2006.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Vitor Martin Bordini, Sébastien Destercke, and Benjamin Quost. Self-learning from pairwise credal labels. In *27th European Conference on Artificial Intelligence, Including 13th Conference on Prestigious Applications of Intelligent Systems (ECAI-PAIS 2024)*, 2024.

Romain Brault, Alex Lambert, Zoltan Szabo, Maxime Sangnier, and Florence d'Alche Buc. Infinite task learning in RKHSs. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1294–1302. PMLR, 16–18 Apr 2019.

Andrea Campagner. Credal learning: Weakly supervised learning from credal sets. In *ECAI 2023*, pp. 327–334. IOS Press, 2023.

Michele Caprio, Maryam Sultana, Eleni G Elia, and Fabio Cuzzolin. Credal learning theory. *Advances in Neural Information Processing Systems*, 37:38665–38694, 2024.

Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pp. 127–135. PMLR, 2025b.

Hao Chen, Ankit Shah, Jindong Wang, Ran Tao, Yidong Wang, Xiang Li, Xing Xie, Masashi Sugiyama, Rita Singh, and Bhiksha Raj. Imprecise label learning: A unified framework for learning with various imprecise label configurations. *Advances in Neural Information Processing Systems*, 37:59621–59654, 2024.

Mengqi Chen, Thomas Berrett, Theodoros Damoulas, and Michele Caprio. Bulk-calibrated credal ambiguity sets: Fast, tractable decision making under out-of-sample contamination. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QO82qIzEsP.

Yaofo Chen. Pytorch cifar models. https://github.com/chenyaofo/pytorch-cifar-models, 2025.

Giorgio Corani and Marco Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(4), 2008.

Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, pp. 49–93, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Sébastien Destercke. Uncertain data in learning: challenges and opportunities. *Conformal and Probabilistic Prediction with Applications*, pp. 322–332, 2022.

Didier Dubois, Henri Prade, and Philippe Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(3):361–377, 1996.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016.

Xabier Gonzalez-Garcia, Siu Lun Chau, Julian Rodemann, Michele Caprio, Krikamol Muandet, Humberto Bustince, Sebastien Destercke, Eyke Hüllermeier, and Yusuf Sale. Quantification of credal uncertainty: A distance-based approach. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 1725–1747. PMLR, 2026.

Romain Guillaume, Inés Couso, and Didier Dubois. Maximum likelihood with coarse data based on robust optimisation. In *Proceedings of the tenth international symposium on imprecise probability: theories and applications*, pp. 169–180. PMLR, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with Bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pp. 822–832. PMLR, 2022.

Paul Hofman, Timo Löhr, Maximilian Muschalik, Yusuf Sale, and Eyke Hüllermeier. Efficient credal prediction through decalibration. In *International Conference on Learning Representations*, volume 2026, pp. 96745–96781, 2026.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Peter J Huber. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518, 1992.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Eyke Hüllermeier, Sébastien Destercke, and Ines Couso. Learning from imprecise data: adjustments of optimistic and pessimistic variants. In *International Conference on Scalable Uncertainty Management*, pp. 266–279. Springer, 2019.

Jens Christian HÜhn and Eyke HÜllermeier. FR3: A fuzzy rule learner for inducing reliable classifiers. *IEEE Transactions on Fuzzy Systems*, 17(1):138–149, 2009. doi: 10.1109/TFUZZ.2008.2005490.

Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1980.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for OOD detection. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Julian Lienen and Eyke Hüllermeier. Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382, 2021a.

Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8583–8591, 2021b.

Julian Lienen, Caglar Demir, and Eyke Hüllermeier. Conformal credal self-supervised learning. In *Conformal and probabilistic prediction with applications*, pp. 214–233. PMLR, 2023.

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.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

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

Denis D Mauá, Fabio G Cozman, Diarmaid Conaty, and Cassio P Campos. Credal sum-product networks. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pp. 205–216. PMLR, 2017.

Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Credal ensembling in multi-class classification. *Machine Learning*, 114(1):19, 2025. ISSN 1573-0565.

Yao Ni, Jeremie Houssineau, Yew-Soon Ong, and Piotr Koniusz. Possibilistic predictive uncertainty for deep learning. In *Forty-third International Conference on Machine Learning*, 2026.

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*, pp. 9617–9626, 2019.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR, 2019.

Côme Rodriguez, Vitor Martin Bordini, Sébastien Destercke, and Benjamin Quost. Self learning using venn-abers predictors. In *Conformal and Probabilistic Prediction with Applications*, pp. 234–250. PMLR, 2023.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 4510–4520. Ieee, 2018.

Lars Schmarje, Claudius Zelenka, Ulf Geisen, Claus-C Glüer, and Reinhard Koch. 2d and 3d segmentation of uncertain local collagen fiber orientations in shg microscopy. In *German Conference on Pattern Recognition*, pp. 374–386. Springer, 2019.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.

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

Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference. In *Proceedings of the Thirty-first Workshop Computational Intelligence*, volume 25, pp. 63, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 45544–45570, 2024.

Anurag Singh, Siu Lun Chau, and Krikamol Muandet. Truthful elicitation of imprecise forecasts. In *Conference on Uncertainty in Artificial Intelligence*, pp. 3898–3919. PMLR, 2025.

Baoye Song, Shihao Zhao, Luyao Dang, Haoguang Wang, and Lin Xu. A survey on learning from data with label noise via deep neural networks. *Systems Science & Control Engineering*, 13(1): 2488120, 2025.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Adane Nega Tarekegn, Mohib Ullah, and Faouzi Alaya Cheikh. Deep learning for multi-label learning: A comprehensive survey. *arXiv preprint arXiv:2401.16549*, 2024.

Varun Venkatesh, Eyke Hüllermeier, Bernd Bischl, and Mina Rezaei. Structured credal learning. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6917–6952. PMLR, 17–21 Aug 2026.

Kaizheng Wang, Fabio Cuzzolin, Shireen Kudukkil Manchingal, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal deep ensembles for uncertainty quantification. In *The Thirty-eighth Annual Conference on 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*, 2025a.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. CreINNs: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, 185:107198, 2025b. ISSN 0893-6080.

Kaizheng Wang, Fabio Cuzzolin, David Moens, and Hans Hallez. Credal ensemble distillation for uncertainty quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26319–26327, 2026a.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. A review of uncertainty representation and quantification in neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2476–2495, 2026b. doi: 10.1109/TPAMI.2025.3626645.

Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin, Siu Lun Chau, David Moens, and Hans Hallez. Learning credal ensembles via distributionally robust optimization. In *Forty-third International Conference on Machine Learning*, 2026c.

Kaizheng Wang, Yunjia Wang, Fabio Cuzzolin, David Moens, Hans Hallez, and Siu Lun Chau. Set-based v.s. distribution-based representations of epistemic uncertainty: A comparative study. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7072–7102. PMLR, 17–21 Aug 2026d.

Frederik Warburg and Marco Miani. Pytorch-laplace. *GitHub. Note: https://github.com/FrederikWarburg/pytorch-laplace*, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Anita Yang, Krikamol Muandet, Michele Caprio, Siu Lun Chau, and Masaki Adachi. Verbalizing LLM's higher-order uncertainty via imprecise probabilities. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 7761–7779. PMLR, 2026.

Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1): 5–21, 2002.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2021.

TABLE OF APPENDIX CONTENT

A Related Work **16**

B Proof **16**

B.1 Proof of Proposition 3.2 16

B.1.1 Setup and Standing Assumptions 16

B.1.2 Reparametrisation by Selections 16

B.1.3 Formal Proof of Proposition 3.2 17

B.2 Proof of Proposition 3.3 18

B.3 Proof of Theorem 4.2 19

B.3.1 Setup and Standing Assumptions 19

B.3.2 Elementary Ingredients for the Proof 20

B.3.3 Proof of the Generic Generalisation Bound 23

B.3.4 Specialisation to the Practical Credal Label Construction 25

C Experiment Implementation Details **26**

C.1 Datasets 26

C.2 Baselines 26

C.3 Training Configurations 27

C.3.1 Annotation Imprecision: Annotator Disagreement 27

C.3.2 Annotation Imprecision: Teacher Predictions 28

C.3.3 Annotation Imprecision: Label Smoothing 28

C.4 Compute Resources 29

C.5 Evaluation Metrics Implementation 29

D Additional Experiment Results **30**

E Evaluation on Statistical Significance Tests **34**

F Comparison with Baselines Trained on Collapsed Labels **38**

G Evaluation with Normalised AUARC **39**

H Evaluation with EU–Prediction Error Correlation **40**

I Ablations on Epistemic Uncertainty Quantification **43**

I.1 Maximum Mean Imprecision 43

I.2 Shannon Entropy Difference 43

J On the Exclusion of OOD Detection for EU Evaluation **45**

A Related Work

Credal sets have attracted increasing attention in the community as a principled framework for representing epistemic uncertainty (EU), with a broad body of work predating their adoption in deep learning and in other statistical applications (Zaffalon, 2002; Corani & Zaffalon, 2008; Corani et al., 2012; Mauá et al., 2017; Shaker & Hüllermeier, 2021; Chau et al., 2025b; Singh et al., 2025; Yang et al., 2026). A central advantage of credal sets lies in their ability to unify set-based and distributional reasoning within a single coherent framework, making them a more natural representation of EU than any single probability distribution (Corani et al., 2012; Hüllermeier & Waegeman, 2021). This is because sets more directly encode ignorance as an absence of knowledge (Dubois et al., 1996): probability distributions, by contrast, carry implicit assumptions that go beyond merely separating plausible from implausible hypotheses (Löhr et al., 2025; Caprio et al., 2024).

In deep learning, credal predictors have been proposed across several paradigms, though all existing approaches assume that training annotations are precise. These include efficient credal predictions via decalibration (Decali) (Hofman et al., 2026), ensemble-based credal predictors (Wang et al., 2024; 2025a; Nguyen et al., 2025; Löhr et al., 2025; Wang et al., 2026a), computationally more intensive credal Bayesian neural networks (Caprio et al., 2024), and credal predictors designed primarily for interval-valued inputs (Wang et al., 2025b). We exclude these methods as baselines for two reasons: (1) they are relatively heavyweight classifiers that conflict with our efficiency objective, and (2) including them would render comparisons unfair given the fundamental difference in computational budget. We also exclude the recent random-set neural network (Manchingal et al., 2025), which outputs belief functions to define credal predictions, as its training requires converting one-hot labels into belief-function form—a procedure incompatible with our experimental setting.

B Proof

B.1 Proof of Proposition 3.2

B.1.1 Setup and Standing Assumptions

Recall that $\mathcal{X}$ is a Polish space with Borel σ-algebra $\mathcal{B}(\mathcal{X})$, $\mathcal{Y} = \{1, \dots, K\}$, and

$$\Delta^{K-1} = \Big\{ p \in \mathbb{R}^K : p_k \geq 0,\ \textstyle\sum_{k=1}^K p_k = 1 \Big\}.$$

Since $\mathcal{X}$ is Polish and $\mathcal{Y}$ is finite, $\mathcal{X} \times \mathcal{Y}$ is Polish. We equip $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with the *weak topology* (the coarsest making $P \mapsto \int f\, dP$ continuous for every $f \in C_b(\mathcal{X} \times \mathcal{Y})$), which is metrisable.

We consider a *Credal Markov Kernel* $Q : \mathcal{X} \rightrightarrows \Delta^{K-1}$ subject to the following.

Assumption 1. $Q(x) \neq \emptyset$ for every $x \in \mathcal{X}$;

Assumption 2. $Q(x)$ is closed and convex for every $x \in \mathcal{X}$;

Assumption 3. Q is *measurable*, i.e., its graph $\mathrm{Gr}(Q) = \{(x, p) : p \in Q(x)\}$ belongs to $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\Delta^{K-1})$.

For a fixed input marginal $P_X \in \mathcal{P}(\mathcal{X})$ and a *Credal Markov Kernel* Q, the induced family of joint laws is defined as follows:

$$\mathcal{M}(P_X, Q) = \big\{ P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : (\pi_{\mathcal{X}})_{\#} P = P_X,\ P_{Y|X=x} \in Q(x),\ P_X\text{-a.e. } x \big\}$$

Under *Assumptions 1–3*, we prove that the set $\mathcal{M}(P_X, Q)$ is nonempty, convex, and weakly compact in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. In particular, it is a *closed convex credal set*. We then show that its admissible risks form a compact interval: the pointwise best-case and worst-case losses can both be realised by joint laws in this set. For this claim, fix a measurable predictor f with measurable per-class losses satisfying $\mathbb{E}_{X \sim P_X}[\max_{k \in \mathcal{Y}} |\ell(f(X), k)|] < \infty$.

B.1.2 Reparametrisation by Selections

Lemma B.1. *For $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with $(\pi_{\mathcal{X}})_{\#} P = P_X$, define the sub-probability measures $\nu_k(A) = P(A \times \{k\})$ for $A \in \mathcal{B}(\mathcal{X})$. Then $\nu_k \ll P_X$, and the Radon–Nikodym derivatives $\eta_k = \frac{d\nu_k}{dP_X} \in [0, 1]$*

satisfy $\sum_{k=1}^{K} \eta_k = 1$ P_X*-a.e. The map*

$$\Phi : P \longmapsto \eta = (\eta_1, \ldots, \eta_K) \in L^1(P_X; \mathbb{R}^K)$$

is an affine bijection from $\{P : (\pi_{\mathcal{X}})_{\#}P = P_X\}$ *onto* $\{\eta \in L^1(P_X; \mathbb{R}^K) : \eta(x) \in \Delta^{K-1}\ P_X$*-a.e.*$\}$*, with conditional* $P_{Y|X=x} = \eta(x)$ *and, for every* $f \in C_b(\mathcal{X} \times \mathcal{Y}) \cong C_b(\mathcal{X})^K$*,*

$$\int_{\mathcal{X}\times\mathcal{Y}} f\, dP = \sum_{k=1}^{K} \int_{\mathcal{X}} f(x,k)\, \eta_k(x)\, dP_X(x). \tag{A1}$$

Consequently, $P \in \mathcal{M}(P_X, Q)$ *if and only if* $\eta \in \mathcal{S}$*, where*

$$\mathcal{S} := \Big\{\eta \in L^1(P_X; \mathbb{R}^K) : \eta(x) \in Q(x)\ P_X\text{-a.e.}\Big\}$$

is the set of integrable selections of Q*.*

Proof. Each ν_k is absolutely continuous w.r.t. P_X because $\nu_k(A) \le P(A \times \mathcal{Y}) = P_X(A)$; the same bound gives $\eta_k \le 1$ a.e., and $\sum_k \nu_k = P_X$ gives $\sum_k \eta_k = 1$ a.e. Finiteness of $\mathcal{Y}$ is precisely what permits the conditional to be a bounded Radon–Nikodym derivative rather than a regular conditional probability. Equation A1 follows by writing $f = (f(\cdot,1), \ldots, f(\cdot,K))$ and integrating each slice against ν_k; it shows P is determined by η through integrals that are affine in η, so Φ is an affine bijection. The final equivalence is immediate from $P_{Y|X=x} = \eta(x)$. □

It therefore suffices to show that $\mathcal{S}$ is nonempty, convex, and weakly compact in L^1, and then transfer the topology along Φ.

B.1.3 Formal Proof of Proposition 3.2

Proof. ***Step 1: nonemptiness.*** By *Assumptions 1–3*, Q is a measurable multifunction with nonempty closed values, so the Kuratowski–Ryll-Nardzewski selection theorem yields a measurable $\eta : \mathcal{X} \to \Delta^{K-1}$ with $\eta(x) \in Q(x)$ for all x. Since η is bounded, $\eta \in \mathcal{S}$; hence $\mathcal{S} \neq \emptyset$ and $\mathcal{M}(P_X, Q) \neq \emptyset$.

Step 2: convexity. Let $\eta, \eta' \in \mathcal{S}$ and $\lambda \in [0,1]$. For a.e. x, we have $\eta(x), \eta'(x) \in Q(x)$, and $Q(x)$ is convex by *Assumption 2*, so $\lambda\,\eta(x) + (1-\lambda)\,\eta'(x) \in Q(x)$. Thus $\mathcal{S}$ is convex; since Φ^{-1} is affine by eq. (A1), $\mathcal{M}(P_X, Q)$ is convex.

Step 3: $\mathcal{S}$ is norm-closed in L^1. Let $\eta^{(m)} \to \eta$ in $L^1(P_X; \mathbb{R}^K)$. Pass to a subsequence with $\eta^{(m_j)}(x) \to \eta(x)$ for a.e. x. For such x, $\eta^{(m_j)}(x) \in Q(x)$ and $Q(x)$ is closed, whence $\eta(x) \in Q(x)$. Therefore $\eta \in \mathcal{S}$.

Step 4: $\mathcal{S}$ is weakly closed in L^1. $\mathcal{S}$ is convex (Step 2) and norm-closed (Step 3). By Mazur's theorem, a convex norm-closed subset of a Banach space is weakly closed. Here the convexity established in Step 2 permits the passage from norm closedness to weak closedness.

Step 5: weak closedness of $\mathcal{M}(P_X, Q)$. Let $P^{(n)} \in \mathcal{M}(P_X, Q)$ with $P^{(n)} \rightharpoonup P$ weakly in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Since $P \mapsto (\pi_{\mathcal{X}})_{\#}P$ is weakly continuous, $(\pi_{\mathcal{X}})_{\#}P = P_X$, so $\eta := \Phi(P)$ and $\eta^{(n)} := \Phi(P^{(n)})$ are well defined. By eq. (A1), for every $g \in C_b(\mathcal{X})$ and each k,

$$\int_{\mathcal{X}} g\, \eta_k^{(n)}\, dP_X \longrightarrow \int_{\mathcal{X}} g\, \eta_k\, dP_X. \tag{A2}$$

The family $(\eta^{(n)})$ takes values in Δ^{K-1}, hence is bounded in L^∞ and uniformly integrable; by the Dunford–Pettis theorem it is relatively weakly compact in L^1. Let ζ be any weak-L^1 cluster point. Then eq. (A2) gives

$$\int g(\zeta_k - \eta_k)\, dP_X = 0$$

for all $g \in C_b(\mathcal{X})$. As $C_b(\mathcal{X})$ is $L^1(P_X)$-determining—for any $A \in \mathcal{B}(\mathcal{X})$, Lusin's theorem provides $g_j \in C_b(\mathcal{X})$ with $\|g_j\|_\infty \le 1$ and $g_j \to \mathbf{1}_A$ in $L^1(P_X)$, whence $\int_A (\zeta_k - \eta_k) dP_X = 0$—we conclude $\zeta = \eta$ a.e. Thus every cluster point equals η, so $\eta^{(n)} \rightharpoonup \eta$ weakly in L^1. Since

$\eta^{(n)} \in \mathcal{S}$ and $\mathcal{S}$ is weakly closed (Step 4), $\eta \in \mathcal{S}$, i.e. $P \in \mathcal{M}(P_X, Q)$. The metrisability of the weak topology in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ makes sequential closedness sufficient.

Step 6: tightness and weak compactness. Fix $\varepsilon > 0$. By Ulam's theorem (P_X is Radon on a Polish space), choose a compact $K_\varepsilon \subseteq \mathcal{X}$ with $P_X(K_\varepsilon) \geq 1 - \varepsilon$. As $\mathcal{Y}$ is finite, $K_\varepsilon \times \mathcal{Y}$ is compact, and for every $P \in \mathcal{M}(P_X, Q)$,

$$P(K_\varepsilon \times \mathcal{Y}) = P_X(K_\varepsilon) \geq 1 - \varepsilon.$$

Hence $\mathcal{M}(P_X, Q)$ is *uniformly tight*, so relatively weakly compact by Prokhorov's theorem. Combined with weak closedness (Step 5), $\mathcal{M}(P_X, Q)$ is weakly compact.

Step 7: the credal risk profile. To identify the full range of admissible risks, we first construct joint laws attaining its endpoints. Write $c_k(x) = \ell(f(x), k)$ and $a(x) = \max_k |c_k(x)|$. The integrability assumption makes a finite P_X-a.e.; if necessary, set all c_k to zero on the common exceptional null set, which does not affect any risk because every admissible joint law has marginal P_X. The function $L(x, q) = \sum_{k=1}^{K} q_k c_k(x)$ is measurable in x, continuous in q, and satisfies $|L(x, q)| \leq a(x)$ for every $q \in \Delta^{K-1}$, up to the same null-set modification.

Since Q is measurable with nonempty compact values, the measurable maximum theorem gives measurable minimum and maximum values and nonempty compact sets of minimisers and maximisers. Measurable selection therefore yields q_+ and q_- such that

$$q_+(x) \in \operatorname*{arg\,min}_{q \in Q(x)} L(x, q), \qquad q_-(x) \in \operatorname*{arg\,max}_{q \in Q(x)} L(x, q).$$

Their induced joint laws, defined by $P_\pm(A \times \{k\}) = \int_A q_{\pm,k}(x)\, dP_X(x)$, belong to $\mathcal{M}(P_X, Q)$. The integrable envelope a ensures that all admissible risks are finite and that the integral representation in Lemma B.1 also applies to these losses. Consequently,

$$\mathbb{E}_{P_+}[\ell(f(X), Y)] = R_+(f), \qquad \mathbb{E}_{P_-}[\ell(f(X), Y)] = R_-(f),$$

with R_+ and R_- as defined in eq. (3). Every $P \in \mathcal{M}(P_X, Q)$ has a conditional selection $\eta(x) \in Q(x)$ P_X-a.e., so integrating the pointwise inequalities $\min_{q \in Q(x)} L(x, q) \leq L(x, \eta(x)) \leq \max_{q \in Q(x)} L(x, q)$ gives

$$R_+(f) \leq \mathbb{E}_P[\ell(f(X), Y)] \leq R_-(f).$$

Conversely, convexity of $\mathcal{M}(P_X, Q)$ ensures that $(1 - t)P_+ + tP_-$ is admissible for each $t \in [0, 1]$, with risk $(1 - t)R_+(f) + tR_-(f)$. Thus every value between the attained endpoints is admissible, proving the compact-interval identity in eq. (2). □

Remark B.2. The compactness argument uses Polishness through metrisability of the weak topology, the determining properties of $C_b(\mathcal{X})$, and tightness of P_X. Finiteness of $\mathcal{Y}$ gives the selection representation, while convexity of the values yields both convexity and weak closedness of the selection set. For a bounded continuous loss on $\mathcal{X} \times \mathcal{Y}$, weak compactness alone ensures that the extreme risks are attained. Step 7 establishes attainment more generally through measurable selection and an integrable envelope; it requires no continuity of the loss in x.

B.2 Proof of Proposition 3.3

Proof. The mixture representation reduces each inner optimisation to allocating the unrestricted probability mass. Fix an input x, write $j = j(x)$ and $\alpha = \alpha(x)$, and consider a prediction p with strictly positive coordinates. Set $c_k = -\log p_k$. Since $\mathcal{Q}_\alpha(x) = \alpha e_j + (1 - \alpha)\Delta^{K-1}$, every admissible target has the form $q = \alpha e_j + (1 - \alpha)r$ for some $r \in \Delta^{K-1}$. This set of mixtures is an ϵ-contamination class with $\epsilon = 1 - \alpha$. Its cross-entropy loss is therefore

$$\ell(p, q) = \alpha c_j + (1 - \alpha) \sum_{k=1}^{K} r_k c_k.$$

The sum is a weighted average of the per-class losses, so it lies between $\min_k c_k$ and $\max_k c_k$. Both bounds are attained by placing all mass of r on a class attaining the corresponding extremum. As $-\log$ is decreasing, the largest loss corresponds to a least probable class, and the smallest loss to a most probable class.

Applying this argument to the two heads, choose $k_{\min} \in \arg\min_k p_{-,k}$ and $k_{\max} \in \arg\max_k p_{+,k}$. An inner maximiser for the pessimistic head and an inner minimiser for the optimistic head are, respectively,

$$q_-^* = \alpha e_j + (1-\alpha)e_{k_{\min}}, \qquad q_+^* = \alpha e_j + (1-\alpha)e_{k_{\max}}.$$

Any tied extremal class gives a valid solution; uniqueness is not required. If a selected class is j, its target reduces to e_j. When $\alpha = 1$, both expressions reduce to the only admissible target e_j. □

B.3 Proof of Theorem 4.2

B.3.1 Setup and Standing Assumptions

The proof connects the joint training objective to the true risk of the averaged predictor. Throughout, Q is a fixed measurable credal kernel satisfying the regularity conditions of Section 3, and we retain the head convention of Section 3.1: the pessimistic head $f_{-,\theta}$ minimises an upper risk, while the optimistic head $f_{+,\theta}$ minimises a lower risk. We use the normalised population and empirical objectives corresponding to eq. (5):

$$J_Q(\theta) := \frac{1}{2}\left\{R_-(f_{-,\theta}) + R_+(f_{+,\theta})\right\}, \quad \widehat{J}_{Q,n}(\theta) := \frac{1}{2}\left\{\widehat{R}_{-,n}(f_{-,\theta}) + \widehat{R}_{+,n}(f_{+,\theta})\right\}. \tag{A3}$$

For $q \in \Delta^{K-1}$ and a probabilistic prediction $f(x) \in \Delta^{K-1}$, the cross-entropy loss is

$$H(q, f(x)) := -\sum_{k=1}^{K} q_k \log f_k(x). \tag{A4}$$

The true risk of any ordinary first-order predictor f under cross-entropy is then

$$R_0(f) := \mathbb{E}_X H(p_0(X), f(X)) = \mathbb{E}_{(X,Y)\sim P_0}[-\log f_Y(X)]. \tag{A5}$$

For clarity, define the pointwise upper and lower losses by

$$\ell_Q^-(f,x) := \max_{q\in Q(x)} H(q, f(x)), \qquad \ell_Q^+(f,x) := \min_{q\in Q(x)} H(q, f(x)). \tag{A6}$$

Their population risks are $R_-(f) = \mathbb{E}_X[\ell_Q^-(f,X)]$ and $R_+(f) = \mathbb{E}_X[\ell_Q^+(f,X)]$, as in Section 3; their empirical counterparts are $\widehat{R}_{\pm,n}(f) = n^{-1}\sum_{i=1}^n \ell_Q^{\pm}(f, X_i)$.

We adopt the following assumptions throughout the proof.

Assumption 1. (Conditional coverage) For P_X-almost every x,

$$p_0(x) \in Q(x). \tag{A7}$$

Assumption 2. (Sampling) The inputs $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_X$, and the corresponding credal labels $Q(X_i)$ are observed. The credal kernel Q and the model class are fixed independently of this sample.

Assumption 3. (Uniformly bounded cross-entropy) There exist $B < \infty$ and a measurable set $\mathcal{X}_0 \subseteq \mathcal{X}$ with $P_X(\mathcal{X}_0) = 1$ such that, simultaneously for every $x \in \mathcal{X}_0$, $\theta \in \mathbf{\Theta}$, and $q \in Q(x)$,

$$0 \le H(q, f_{-,\theta}(x)) \le B, \qquad 0 \le H(q, f_{+,\theta}(x)) \le B. \tag{A8}$$

The common set $\mathcal{X}_0$ ensures that this bound also applies to parameters selected from the observed sample.

Assumption 4. (Bounded variation in classwise losses) The difference between the largest and smallest classwise log losses, called their oscillation, is uniformly bounded: there exists $L < \infty$ such that every prediction used by the pair class satisfies

$$\operatorname{osc}(-\log f(x)) := \max_k[-\log f_k(x)] - \min_k[-\log f_k(x)] \le L \tag{A9}$$

for every $x \in \mathcal{X}_0$, simultaneously for all parameters and both heads.

Assumption 5. (Approximate joint ERM) The learned parameter $\widehat{\theta}_n$ satisfies

$$\widehat{J}_{Q,n}(\widehat{\theta}_n) \le \inf_{\theta\in\mathbf{\Theta}} \widehat{J}_{Q,n}(\theta) + \eta_n, \tag{A10}$$

where $\eta_n \ge 0$ is the optimisation error in the normalised objective. If the unnormalised objective is optimised to tolerance $\widetilde{\eta}_n$, then $\eta_n = \widetilde{\eta}_n/2$.

Assumption 6. (Nonempty comparator class) At least one ordinary predictor can be represented by making the two heads identical. Thus, the corresponding comparator class is nonempty:

$$\mathcal{F}_\Delta := \left\{ f : \exists \theta_f \in \mathbf{\Theta} \text{ such that } f_{-,\theta_f} = f_{+,\theta_f} = f \right\} \neq \varnothing. \tag{A11}$$

Let $\widehat{\theta}_n$ be an η_n-approximate empirical minimiser of eq. (A3), and let

$$\widehat{\bar{f}}_n := \bar{f}_{\widehat{\theta}_n} = \frac{f_{-,\widehat{\theta}_n} + f_{+,\widehat{\theta}_n}}{2}$$

denote the corresponding averaged predictor defined in Section 3.1. The learned pair need not lie on the diagonal $f_+ = f_-$, so we compare its true risk with that of the best ordinary predictor in $\mathcal{F}_\Delta$.

Under assumptions in eqs. (A7) to (A11), we prove the generic generalisation bound of POCC stated in Theorem 4.2:

$$R_0(\widehat{\bar{f}}_n) - \inf_{f\in\mathcal{F}_\Delta} R_0(f) \le \underbrace{4\mathfrak{R}_n(\mathcal{J}_Q)}_{\text{model class complexity}} + \underbrace{2B\sqrt{\frac{\log(2/\delta)}{2n}}}_{\text{finite-sample estimation error}} + \underbrace{\eta_n}_{\text{optimisation gap}} + \underbrace{L\bar{d}_Q}_{\text{annotation imprecision}}.$$

B.3.2 Elementary Ingredients for the Proof

We first establish several elementary results used in the proof of Theorem 4.2. Coverage makes the true conditional distribution feasible in both pointwise optimisations, so $R_+(f) \le R_0(f) \le R_-(f)$. The following argument controls how far either endpoint can lie from the true risk.

Credal Diameter Controls Ambiguity Gaps. Define the pointwise total-variation (TV) diameter

$$d_Q(x) := \sup_{q,r\in Q(x)} \mathrm{TV}(q,r), \qquad \bar{d}_Q := \mathbb{E}_X[d_Q(X)], \tag{A12}$$

where, on the finite label space,

$$\mathrm{TV}(q,r) = \frac{1}{2}\|q-r\|_1.$$

Lemma B.3 (Total Variation–Oscillation Inequality)**.** *For any $q, r \in \Delta^{K-1}$ and $a \in \mathbb{R}^K$,*

$$|\langle q-r, a\rangle| \le \mathrm{TV}(q,r)\ \mathrm{osc}(a), \qquad \mathrm{osc}(a) = \max_k a_k - \min_k a_k. \tag{A13}$$

Proof. Let

$$m := \frac{\max_k a_k + \min_k a_k}{2}.$$

Because q and r both sum to one,

$$\langle q-r, \mathbf{1}\rangle = 0.$$

Consequently,

$$\langle q-r, a\rangle = \langle q-r, a-m\mathbf{1}\rangle.$$

By Hölder's inequality,

$$|\langle q-r, a-m\mathbf{1}\rangle| \le \|q-r\|_1 \|a-m\mathbf{1}\|_\infty.$$

The two factors satisfy

$$\|q-r\|_1 = 2\,\mathrm{TV}(q,r), \qquad \|a-m\mathbf{1}\|_\infty = \frac{\mathrm{osc}(a)}{2}.$$

Multiplying them gives eq. (A13). □

The upper and lower ambiguity gaps measure the two endpoint deviations from the true risk:

$$A_Q^-(f) := R_-(f) - R_0(f) \geq 0, \quad A_Q^+(f) := R_0(f) - R_+(f) \geq 0. \tag{A14}$$

Proposition B.4 (Generic Ambiguity-gap Bounds)**.** *Under conditional coverage, for every f,*

$$0 \leq A_Q^-(f) \leq \mathbb{E}_X[d_Q(X)\ \mathrm{osc}(-\log f(X))]\,, \tag{A15}$$

$$0 \leq A_Q^+(f) \leq \mathbb{E}_X[d_Q(X)\ \mathrm{osc}(-\log f(X))]\,. \tag{A16}$$

Under the uniform oscillation bound eq. (A9)*, both gaps are at most $L\bar{d}_Q$.*

Proof. Set

$$a_f(x) := -\log f(x) = \big(-\log f_1(x), \ldots, -\log f_K(x)\big).$$

For the upper gap, use coverage and write

$$\ell_Q^-(f,x) - H(p_0(x), f(x)) = \sup_{q \in Q(x)} q^\top a_f(x) - p_0(x)^\top a_f(x) = \sup_{q \in Q(x)} (q - p_0(x))^\top a_f(x).$$

For each feasible q, both q and $p_0(x)$ belong to $Q(x)$, so

$$\mathrm{TV}(q, p_0(x)) \leq d_Q(x).$$

Applying eq. (A13),

$$(q - p_0(x))^\top a_f(x) \leq d_Q(x)\,\mathrm{osc}(a_f(x)).$$

The right-hand side no longer depends on q, so the same bound holds after taking the supremum. Integrating proves eq. (A15).

For the lower gap,

$$\begin{aligned} H(p_0(x), f(x)) - \ell_Q^+(f,x) &= p_0(x)^\top a_f(x) - \inf_{q \in Q(x)} q^\top a_f(x) \\ &= \sup_{q \in Q(x)} (p_0(x) - q)^\top a_f(x), \end{aligned}$$

and the identical argument proves eq. (A16). The final assertion follows directly from $\mathrm{osc}(-\log f(X)) \leq L$ in eq. (A9). □

Population Risk of an Arbitrary Averaged Dual-head Predictor. Next, we establish the central bridge between the algorithm's joint credal objective and the true risk of its averaged prediction.

Proposition B.5 (Population Risk Bound for the Average)**.** *For every $\theta \in \mathbf{\Theta}$, under conditional coverage,*

$$R_0(\bar{f}_\theta) \leq J_Q(\theta) + \frac{1}{2}\left\{A_Q^+(f_{+,\theta}) - A_Q^-(f_{-,\theta})\right\}. \tag{A17}$$

Consequently,

$$R_0(\bar{f}_\theta) \leq J_Q(\theta) + \frac{1}{2}A_Q^+(f_{+,\theta}), \tag{A18}$$

$$\leq J_Q(\theta) + \frac{1}{2}\mathbb{E}_X[d_Q(X)\,\mathrm{osc}(-\log f_{+,\theta}(X))]\,, \tag{A19}$$

$$\leq J_Q(\theta) + \frac{L}{2}\bar{d}_Q. \tag{A20}$$

Proof. Start with convexity of the true risk:

$$R_0(\bar{f}_\theta) \leq \frac{1}{2}R_0(f_{-,\theta}) + \frac{1}{2}R_0(f_{+,\theta}). \tag{A21}$$

Use the definitions of the ambiguity gaps in eq. (A14):

$$R_0(f_{-,\theta}) = R_-(f_{-,\theta}) - A_Q^-(f_{-,\theta}), \quad R_0(f_{+,\theta}) = R_+(f_{+,\theta}) + A_Q^+(f_{+,\theta}).$$

Substitute both identities into eq. (A21):

$$\begin{aligned} R_0(\bar{f}_\theta) &\le \frac{1}{2}\big[R_-(f_{-,\theta}) - A_Q^-(f_{-,\theta})\big] + \frac{1}{2}\big[R_+(f_{+,\theta}) + A_Q^+(f_{+,\theta})\big] \\ &= \frac{1}{2}\big[R_-(f_{-,\theta}) + R_+(f_{+,\theta})\big] + \frac{1}{2}\big[A_Q^+(f_{+,\theta}) - A_Q^-(f_{-,\theta})\big] \\ &= J_Q(\theta) + \frac{1}{2}\big[A_Q^+(f_{+,\theta}) - A_Q^-(f_{-,\theta})\big]. \end{aligned}$$

This proves eq. (A17). Because $A_Q^-(f_{-,\theta}) \ge 0$, dropping the negative term proves eq. (A18). Applying eq. (A16) to the optimistic head gives eq. (A19). Finally, the uniform oscillation bound gives eq. (A20). □

Uniform Convergence for the Joint Infinite Hypothesis Class. Define

$$j_\theta(x) := \frac{1}{2}\left\{\ell_Q^-(f_{-,\theta}, x) + \ell_Q^+(f_{+,\theta}, x)\right\}, \tag{A22}$$

and the induced real-valued class

$$\mathcal{J}_Q := \{j_\theta : \theta \in \mathbf{\Theta}\}. \tag{A23}$$

Then

$$J_Q(\theta) = \mathbb{E}_X[j_\theta(X)], \qquad \widehat{J}_{Q,n}(\theta) = \frac{1}{n}\sum\nolimits_{i=1}^{n} j_\theta(X_i).$$

By the common bound in eq. (A8), every j_θ takes values in $[0, B]$ on the same set $\mathcal{X}_0$ simultaneously for every $\theta \in \Theta$.

For a real-valued class $\mathcal{G}$, define the expected absolute Rademacher complexity

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_{X_{1:n},\sigma_{1:n}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right|\right], \tag{A24}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher signs.

Theorem B.6 (Uniform Convergence of the Joint Objective)**.** *Under i.i.d. sampling and the uniform boundedness assumption eq.* (A8)*, with probability at least* $1 - \delta$*,*

$$\sup_{\theta\in\mathbf{\Theta}}\left|J_Q(\theta) - \widehat{J}_{Q,n}(\theta)\right| \le \varepsilon_{n,\mathrm{joint}}(\delta), \tag{A25}$$

where

$$\varepsilon_{n,\mathrm{joint}}(\delta) := 2\mathfrak{R}_n(\mathcal{J}_Q) + B\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{A26}$$

No finiteness or countability of $\mathbf{\Theta}$ *is required, provided the relevant suprema are measurable and the displayed Rademacher complexity is well defined.*

Proof. Write $Pj := \mathbb{E}[j(X)]$ and $P_n j := n^{-1}\sum_{i=1}^{n} j(X_i)$. Define the random uniform deviation

$$\Phi(X_{1:n}) := \sup_{j\in\mathcal{J}_Q}|Pj - P_n j|.$$

We prove the result in two stages.

Stage 1: symmetrisation. Let $X_1', \ldots, X_n'$ be an independent ghost sample from P_X, and let

$$P_n' j := \frac{1}{n}\sum_{i=1}^{n} j(X_i').$$

Since $Pj = \mathbb{E}_{X'_{1:n}}[P_n' j]$, Jensen's inequality gives

$$\begin{aligned} \mathbb{E}_{X_{1:n}}\Phi &= \mathbb{E}_{X_{1:n}}\sup_{j\in\mathcal{J}_Q}\left|\mathbb{E}_{X'_{1:n}}[P_n' j] - P_n j\right| \\ &\le \mathbb{E}_{X_{1:n},X'_{1:n}}\sup_{j\in\mathcal{J}_Q}|P_n' j - P_n j| \\ &= \mathbb{E}_{X,X'}\sup_{j\in\mathcal{J}_Q}\left|\frac{1}{n}\sum_{i=1}^{n}\big(j(X_i') - j(X_i)\big)\right|. \end{aligned}$$

For each i, the pair (X_i, X_i') is exchangeable. Introducing independent Rademacher signs therefore leaves the joint distribution of the difference unchanged:

$$\mathbb{E}\Phi \le \mathbb{E}_{X,X',\sigma} \sup_{j\in\mathcal{J}_Q} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \big(j(X_i') - j(X_i)\big) \right|.$$

Use the triangle inequality inside the supremum:

$$\sup_{j\in\mathcal{J}_Q} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \big(j(X_i') - j(X_i)\big) \right| \quad \le \sup_{j\in\mathcal{J}_Q} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i j(X_i') \right| + \sup_{j\in\mathcal{J}_Q} \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i j(X_i) \right|.$$

Both terms have the same expectation. Hence

$$\mathbb{E}\Phi \le 2\mathfrak{R}_n(\mathcal{J}_Q). \tag{A27}$$

Stage 2: concentration around the expectation. Since $P_X(\mathcal{X}_0) = 1$, we may regard the inputs as taking values in $\mathcal{X}_0$. By assumption, the bound B holds simultaneously for every $\theta \in \Theta$, both heads, and every admissible target $q \in \mathcal{Q}(x)$ on this common set, i.e.,

$$\sup_{\theta\in\Theta} \sup_{q\in\mathcal{Q}(x)} \ell\big(p_{\pm,\theta}(x), q\big) \le B \qquad \text{for all } x \in \mathcal{X}_0.$$

Consequently, every $j \in \mathcal{J}_Q$ is bounded in $[0, B]$ on $\mathcal{X}_0^n$ simultaneously across all parameters, which is the bounded-differences hypothesis required by McDiarmid's inequality. On $\mathcal{X}_0^n$, replacing one observation X_i by any $\widetilde{X}_i \in \mathcal{X}_0$ therefore changes $P_n j$ by at most B/n:

$$|P_n j - P_n^{(i)} j| \le \frac{B}{n}.$$

Taking a supremum cannot increase the replacement sensitivity beyond B/n, so

$$|\Phi(X_{1:n}) - \Phi(X_1, \ldots, \widetilde{X}_i, \ldots, X_n)| \le \frac{B}{n}.$$

McDiarmid's inequality therefore yields

$$\mathbb{P}\{\Phi - \mathbb{E}\Phi \ge t\} \le \exp\left(-\frac{2nt^2}{B^2}\right).$$

Choose

$$t = B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Then, with probability at least $1 - \delta$,

$$\Phi \le \mathbb{E}\Phi + B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Combining this with eq. (A27) proves eq. (A25)– eq. (A26). □

B.3.3 Proof of the Generic Generalisation Bound

Step 1. We first combine the population relation in eq. (A20) with uniform convergence and approximate empirical risk minimisation.

Theorem B.7 (Risk Bound Relative to the Best Joint Credal Objective)**.** *Under the assumptions in eqs.* (A7) *to* (A10)*, with probability at least* $1 - \delta$*,*

$$R_0(\widehat{\widehat{f}}_n) \le \inf_{\theta\in\mathbf{\Theta}} J_Q(\theta) + 2\varepsilon_{n,\mathrm{joint}}(\delta) + \eta_n + \frac{L}{2}\overline{d}_Q. \tag{A28}$$

Proof. Work on the event eq. (A25). We proceed one line at a time.

First, apply eq. (A20) to the learned pair:

$$R_0(\widehat{\widehat{f}}_n) \le J_Q(\widehat{\theta}_n) + \frac{L}{2}\overline{d}_Q. \tag{A29}$$

Uniform convergence at $\widehat{\theta}_n$ gives

$$J_Q(\widehat{\theta}_n) \le \widehat{J}_{Q,n}(\widehat{\theta}_n) + \varepsilon_{n,\text{joint}}. \tag{A30}$$

Approximate empirical minimisation gives

$$\widehat{J}_{Q,n}(\widehat{\theta}_n) \le \inf_{\theta\in\mathbf{\Theta}} \widehat{J}_{Q,n}(\theta) + \eta_n. \tag{A31}$$

Uniform convergence also gives, simultaneously for every comparator θ,

$$\widehat{J}_{Q,n}(\theta) \le J_Q(\theta) + \varepsilon_{n,\text{joint}}.$$

Taking the infimum over θ,

$$\inf_{\theta\in\mathbf{\Theta}} \widehat{J}_{Q,n}(\theta) \le \inf_{\theta\in\mathbf{\Theta}} J_Q(\theta) + \varepsilon_{n,\text{joint}}. \tag{A32}$$

Combining eq. (A29)– eq. (A32),

$$\begin{aligned} R_0(\widehat{\widehat{f}}_n) &\le J_Q(\widehat{\theta}_n) + \frac{L}{2}\overline{d}_Q \\ &\le \widehat{J}_{Q,n}(\widehat{\theta}_n) + \varepsilon_{n,\text{joint}} + \frac{L}{2}\overline{d}_Q \\ &\le \inf_\theta \widehat{J}_{Q,n}(\theta) + \eta_n + \varepsilon_{n,\text{joint}} + \frac{L}{2}\overline{d}_Q \\ &\le \inf_\theta J_Q(\theta) + 2\varepsilon_{n,\text{joint}} + \eta_n + \frac{L}{2}\overline{d}_Q. \end{aligned}$$

This is eq. (A28). □

Step 2. Theorem B.7 compares against the best two-head credal objective. We now convert it into an excess true-risk bound.

Lemma B.8 (A Diagonal Pair is Close to Its True Risk)**.** *For every $f \in \mathcal{F}_\Delta$,*

$$\frac{1}{2}\big[R_-(f) + R_+(f)\big] \le R_0(f) + \frac{L}{2}\overline{d}_Q. \tag{A33}$$

Proof. Using the ambiguity gaps,

$$\begin{aligned} \frac{1}{2}\big[R_-(f) + R_+(f)\big] - R_0(f) &= \frac{1}{2}\big[A_Q^-(f) - A_Q^+(f)\big] \\ &\le \frac{1}{2}A_Q^-(f) \\ &\le \frac{1}{2}\mathbb{E}_X[d_Q(X)\,\mathrm{osc}(-\log f(X))] \\ &\le \frac{L}{2}\overline{d}_Q. \end{aligned}$$

This proves eq. (A33). □

Step 3. We are now ready to complete the proof of Theorem 4.2:

$$\boxed{R_0(\widehat{\widehat{f}}_n) - \inf_{f\in\mathcal{F}_\Delta} R_0(f) \le 4\mathfrak{R}_n(\mathcal{J}_Q) + 2B\sqrt{\frac{\log(2/\delta)}{2n}} + \eta_n + L\overline{d}_Q} \tag{A34}$$

Proof. For every $f \in \mathcal{F}_\Delta$, by definition there exists a parameter θ_f with

$$f_{-,\theta_f} = f_{+,\theta_f} = f.$$

Therefore

$$J_Q(\theta_f) = \frac{1}{2}\big[R_-(f) + R_+(f)\big].$$

By eq. (A33),

$$J_Q(\theta_f) \le R_0(f) + \frac{L}{2}\overline{d}_Q.$$

Taking the infimum over $f \in \mathcal{F}_\Delta$ and noting that the infimum over all θ is no larger than the infimum over diagonal parameters,

$$\inf_{\theta\in\mathbf{\Theta}} J_Q(\theta) \le \inf_{f\in\mathcal{F}_\Delta} R_0(f) + \frac{L}{2}\overline{d}_Q. \tag{A35}$$

Substitute eq. (A35) into eq. (A28):

$$\begin{aligned} R_0(\widehat{\widehat{f}}_n) &\le \inf_{f\in\mathcal{F}_\Delta} R_0(f) + \frac{L}{2}\overline{d}_Q + 2\varepsilon_{n,\mathrm{joint}} + \eta_n + \frac{L}{2}\overline{d}_Q \\ &= \inf_{f\in\mathcal{F}_\Delta} R_0(f) + 2\varepsilon_{n,\mathrm{joint}} + \eta_n + L\overline{d}_Q. \end{aligned}$$

Subtract $\inf_{f\in\mathcal{F}_\Delta} R_0(f)$ and expand

$$2\varepsilon_{n,\mathrm{joint}}(\delta) = 4\mathfrak{R}_n(\mathcal{J}_Q) + 2B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

This gives eq. (A34). □

B.3.4 Specialisation to the Practical Credal Label Construction

For the credal construction in eq. (8), i.e.,

$$\mathcal{Q}_\alpha(x) = \{q \in \Delta^{K-1} : q_{j(x)} \ge \alpha(x)\},$$

the conditional coverage becomes the transparent condition

$$p_{0,j(x)}(x) \ge \alpha(x) \qquad P_X\text{-a.s.} \tag{A36}$$

Suppose each head is a softmax of logits $s_{\pm,\theta}(x)$ satisfying, simultaneously for every $\theta \in \mathbf{\Theta}$, both heads, and every x in a common measurable set $\mathcal{X}_0$ with $P_X(\mathcal{X}_0) = 1$,

$$\|s_{\pm,\theta}(x)\|_\infty \le A. \tag{A37}$$

Then

$$f_{\pm,\theta,k}(x) \ge \frac{e^{-2A}}{K}, \qquad B = \log K + 2A. \tag{A38}$$

Moreover,

$$\mathrm{osc}(-\log f_{\pm,\theta}(x)) = \max_k s_{\pm,\theta,k}(x) - \min_k s_{\pm,\theta,k}(x) \le 2A, \tag{A39}$$

so $L = 2A$.

Verification. For the probability lower bound,

$$f_k(x) = \frac{e^{s_k(x)}}{\sum_{\ell=1}^K e^{s_\ell(x)}} \ge \frac{e^{-A}}{Ke^A} = \frac{e^{-2A}}{K}.$$

Thus $-\log f_k(x) \le \log K + 2A$, and any cross-entropy, being a convex combination of these coordinate losses, is bounded by the same B.

Also,

$$-\log f_k(x) = \log\left(\sum\nolimits_{\ell=1}^K e^{s_\ell(x)}\right) - s_k(x).$$

The common log-sum-exp term cancels when taking the oscillation, leaving $\max_k s_k - \min_k s_k \le 2A$. □

Applying the generic bound eq. (A34) with $B = \log K + 2A$, $L = 2A$, and

$$\overline{d}_Q = \mathbb{E}_X[d_{Q_{\alpha,j}}(X)] = \mathbb{E}_X[1 - \alpha(X)],$$

we obtain the specialised bound:

$$\boxed{R_0(\widehat{\widehat{f}}_n) - \inf_{f\in\mathcal{F}_\Delta} R_0(f) \le 4\mathfrak{R}_n(\mathcal{J}_{Q_{\alpha,j}}) + 2(\log K + 2A)\sqrt{\frac{\log(2/\delta)}{2n}} + \eta_n + 2A\,\mathbb{E}[1 - \alpha(X)].}$$

C Experiment Implementation Details

This section details the experimental setup for our experiments. The datasets are drawn from existing literature. *To further support reproducibility, the core implementation codes are included in the supplementary materials and will be released publicly upon acceptance under a licence permitting free use for research purposes.*

C.1 Datasets

CIFAR-10. The CIFAR-10 dataset (Krizhevsky, 2009) consists of 60,000 colour images (32×32) across 10 classes, split into 50,000 training and 10,000 test images.

CIFAR-10H. The CIFAR-10H (Peterson et al., 2019) provides human soft labels for the CIFAR-10 test set, with 511400 annotations from 2571 workers and approximately 51 labels per image, capturing human perceptual variability.

CIFAR-100. The CIFAR-100 dataset (Krizhevsky, 2009) contains 60,000 colour images (32×32) across 100 fine-grained classes, grouped into 20 superclasses, with 600 images per class and a 50,000/10,000 train/test split.

MiceBone. The MiceBone dataset (Schmarje et al., 2022) comprises 2D slices from 3D scans of mouse bone collagen fibres, originally introduced in (Schmarje et al., 2019). It consists of three classes: similar ("g") and dissimilar ("ug") collagen fibre orientations, and not relevant ("nr") regions due to noise or background. Annotations were collected from four annotators (three hired workers and one domain expert), who received reference-guided training before annotation. An initial familiarisation phase was discarded, and the final labels are based on majority vote.

TreeVersity. The Treeversity#1 dataset (Schmarje et al., 2022) originates from a crowdsourced plant image collection of Arnold Arboretum of Harvard University with up to 22 original tags, simplified into six unified classes. Images were cropped to remove metadata. To capture annotator disagreement, only images with at least three annotations are retained; annotations from different users often yield conflicting tags for the same image. The total number of annotations and the agreement with the majority vote per image are detailed in (Schmarje et al., 2022).

C.2 Baselines

Evidential Deep Learning (EDL). Our implementation of EDL (Sensoy et al., 2018) follows Hofman et al. (2026). A single neural network is employed with a SoftPlus activation on the final layer to ensure non-negative evidence outputs. The model is trained using the Type II maximum likelihood loss with a Kullback–Leibler divergence regularisation term, whose coefficient follows a warm-up schedule $\lambda_i = \min(1, i/10)$ at epoch i.

At inference time, the predicted evidence parameterises a Dirichlet distribution. For an input x, let $e(x) = [e_1(x), \ldots, e_K(x)] \in \mathbb{R}^K_{\geq 0}$ denotes the output evidence, where K is the number of classes. The Dirichlet parameters and total strength are then given by

$$\alpha(x) = e(x) + \mathbf{1}, \qquad S(x) = \sum\nolimits_{k=1}^{K} \alpha_k(x). \tag{A40}$$

Epistemic uncertainty (EU) is quantified as follows (Sensoy et al., 2018):

$$\mathrm{EU}_{\mathrm{EDL}}(x) = \frac{K}{S(x)} = \frac{K}{\sum_{k=1}^{K}\big(e_k(x)+1\big)}. \tag{A41}$$

Laplace Bridge with Bayesian Deep Networks (LbBnn). LbBnn (Hobbhahn et al., 2022) starts from a standard neural network and uses the Laplace Bridge to approximate a Bayesian posterior efficiently. It builds a Gaussian approximation around the network's logits and then applies the Laplace Bridge to analytically map it to a Dirichlet distribution over softmax outputs. This closed-form approximation eliminates the need for sampling, yielding efficient uncertainty estimates. Given the Dirichlet prediction, the EU is quantified similarly to EDL via eq. (A41). Our implementation is based on Warburg & Miani (2023), adopting the last-layer and Kronecker-factored approximation for Laplace approximation to balance computational complexity and model performance.

Efficient Credal Predictions through Decalibration (Decali). Decali (Hofman et al., 2026) is implemented as a post-hoc method following the procedure described in the original work. Given the logits of a maximum likelihood estimation predictor, Decali solves two convex optimisation problems for each class to determine the lower and upper bounds of a plausible probability interval under a relative likelihood constraint, referred to as an α-cut. The relative likelihood budget is estimated from the training data. Each optimisation problem perturbs the logit of a single class by adding a constant, thereby obtaining the corresponding lower or upper probability bound. We set $\alpha = 1.0$, which was reported to achieve the best epistemic uncertainty quantification performance in downstream tasks (Hofman et al., 2026).

At inference time, given an input x, let $\left[\underline{p}_k(x), \overline{p}_k(x)\right]$ denote the resulting class-wise probability intervals. The EU is then quantified as (Wang et al., 2024):

$$\mathrm{EU}_{\mathrm{Decali}}(x) = \max_{p \in C(x)} \sum\nolimits_{k=1}^{K} -p_k \log_2 p_k - \min_{p \in C(x)} \sum\nolimits_{k=1}^{K} -p_k \log_2 p_k, \tag{A42}$$

where

$$C(x) = \left\{ p \in \Delta^{K-1} : \underline{p}_k(x) \leq p_k \leq \overline{p}_k(x),\ k = 1, \ldots, K \right\}.$$

Test accuracy and ECE are evaluated using the original softmax probabilities before perturbation (Hofman et al., 2026).

Dirichlet-Approximated Possibilistic Posterior Predictions (DAPPr). DAPPr (Ni et al., 2026) defines a possibilistic posterior over model parameters, projects it to the prediction space via supremum operators to obtain an epistemic uncertainty function, and approximates this projection using learned Dirichlet possibility functions. This approximation yields a closed-form solution under the high-capacity assumption of modern overparameterised networks with cross-entropy loss, resulting in a simple training objective with minimal regularisation. We implement DAPPr in our evaluation by following the original work. At inference time, DAPPr outputs a Dirichlet distribution, with EU quantified via the same formulation as EDL in eq. (A41).

Monte Carlo Dropout (MCDO). MCDO (Gal & Ghahramani, 2016) approximates Bayesian inference by retaining dropout during inference. Multiple stochastic forward passes are performed, each sampling a different sub-network via distinct dropout masks. Following the practice, we set the dropout rate to 0.3 and use $M = 5$ forward passes. Epistemic uncertainty is quantified using the approximate mutual information (MI):

$$\mathrm{EU}_{\mathrm{MCDO}}(x) = \sum\nolimits_{k=1}^{K} -\tilde{p}_k(x) \log_2 \tilde{p}_k(x) - \frac{1}{M} \sum\nolimits_{i=1}^{M} \sum\nolimits_{k=1}^{K} -p_{i,k}(x) \log_2 p_{i,k}(x), \tag{A43}$$

where $\tilde{p}_k$ is the k-th component of the averaged prediction $\tilde{p} = \frac{1}{M} \sum_{i=1}^{M} p_i$. This MI corresponds to the standard decomposition of total uncertainty into aleatoric and epistemic components (Wang et al., 2026b).

Deep Ensembles (DE). DE (Lakshminarayanan et al., 2017) quantifies prediction uncertainty by training an ensemble of M neural networks independently with different random initialisations. At inference time, each ensemble member produces a probability vector; the ensemble prediction is obtained by averaging these outputs. Epistemic uncertainty is quantified via the approximate mutual information as in eq. (A43), where the M sampled probability vectors correspond to the outputs of the M ensemble members. In our experiment, $M = 5$.

C.3 Training Configurations

C.3.1 Annotation Imprecision: Annotator Disagreement

For the **CIFAR-10H** experiments, all methods are trained from scratch on the CIFAR-10H dataset using a ResNet-18 (He et al., 2016) backbone. Evaluation is performed on the last 10,000 images of the CIFAR-10 training set (indices 40,000–50,000 under standard torchvision ordering, without shuffling), which do not overlap with the CIFAR-10H data drawn from the original CIFAR-10 test set. For the **MiceBone** and **TreeVersity** datasets, all methods adopt an 80%:20% train–test split and use a DenseNet-121 (Huang et al., 2017) backbone initialised with ImageNet pretrained weights. For these two datasets, MCDO and LbBnn are excluded due to the lack of a compatible pre-trained model and out-of-memory errors on an NVIDIA RTX 5090, respectively. All methods are trained

with cross-entropy loss using soft labels. Comprehensive hyperparameter configurations are provided in Table A1.

Table A1: Hyperparameters used for annotator-disagreement experiments.

Hyperparameter	**CIFAR-10H**	**MiceBone**	**TreeVersity**
Backbone	ResNet-18	DenseNet-121	DenseNet-121
Classes	10	3	6
Input size	$(3 \times 32 \times 32)$	$(3 \times 224 \times 224)$	$(3 \times 224 \times 224)$
Batch size	128	128	128
Epochs	150	60	60
Learning rate	0.1	0.00001	0.00001
Weight decay	0.0005	0.0005	0.0005
Optimiser	SGD	Adam	Adam
Learning rate scheduler	Cosine Annealing	–	–

C.3.2 Annotation Imprecision: Teacher Predictions

In this section, we use predictions from a teacher model (Hinton et al., 2015) on the training data to simulate annotation imprecision. Specifically, given an input x, let $h_{\text{teacher}}(\cdot)$ denote a well-trained teacher model, and let $z(x) = h_{\text{teacher}}(x)$ denotes its logit output. The soft label distribution q_L is then obtained via the standard softmax with temperature:

$$q_{L,k}(x) = \frac{\exp\big(z_k(x)/T\big)}{\sum_{j=1}^{K} \exp\big(z_j(x)/T\big)}, \quad k = 1, \ldots, K, \tag{A44}$$

where T is the temperature parameter. We experiment with temperatures $T \in \{1.0, 2.5\}$ and three publicly available teacher backbones (Chen, 2025) (ResNet-20, ShuffleNetV2-0.5× (Ma et al., 2018), and MobileNetV2-x0-5 (Sandler et al., 2018)) for both CIFAR-10 and CIFAR-100 experiments.

An 80%:20% train–test split is adopted for all experiments. For **CIFAR-10**, all methods are trained from scratch using a ResNet-18 backbone with teacher-predicted labels, and a cosine annealing learning rate scheduler is applied. For **CIFAR-100**, all methods are trained from scratch using a WRN-28-4 backbone with teacher-predicted labels, and a learning rate schedule with linear warmup and cosine annealing is employed: the learning rate is linearly ramped up from 0.001 to the base learning rate during the first 5 epochs, and then gradually decays to 10^{-6} following a cosine curve for the remaining training. Detailed hyperparameter configurations are provided in Table A2.

Table A2: Hyperparameters used for both teacher-prediction and label-smoothing experiments.

Hyperparameter	**CIFAR-10**	**CIFAR-100**
Backbone	ResNet-18	WRN-28-4
Classes	10	100
Input size	$(3 \times 32 \times 32)$	$(3 \times 32 \times 32)$
Batch size	128	128
Epochs	200	200
Learning rate	0.1	0.1
Weight decay	0.0005	0.0005
Optimiser	SGD	SGD
Learning rate scheduler	Cosine Annealing	Warmup Cosine Annealing

C.3.3 Annotation Imprecision: Label Smoothing

In this section, we adopt label smoothing (Szegedy et al., 2016) to simulate annotation imprecision in the training data. Given a hard one-hot label e_y, the soft label distribution q_L is constructed as:

$$q_{L,k} = (1 - \epsilon)\, \mathbf{1}_{k=y} + \frac{\epsilon}{K}, \tag{A45}$$

where ϵ is the smoothing coefficient, varied over $\{0.05, 0.1, 0.15, 0.2\}$ in our experiments, and $\mathbf{1}_{k=y}$ is the indicator function. The resulting credal label for supervision is then obtained via eq. (8).

For the **CIFAR-10** experiments, all methods are trained from scratch on the CIFAR-10 dataset using a ResNet-18 backbone with label smoothing. We apply cosine annealing as the learning rate scheduler. For the **CIFAR-100** experiments, all methods are trained from scratch on the CIFAR-100 dataset using a WRN-28-4 backbone with label smoothing. We employ a learning rate schedule with linear warmup and cosine annealing: the learning rate is linearly ramped up from 0.001 to the base learning rate during the first 5 epochs, and then gradually decays to 10^{-6} following a cosine curve for the remaining training. An 80%:20% train–test split is adopted for all experiments. Comprehensive hyperparameter configurations are provided in Table A2.

C.4 Compute Resources

All experiments in this work are conducted using the computing resources listed in Table A3.

Table A3: Specifications of Computing Resources

Component	Specification
CPU	Intel Core Ultra 24C/24T, 5.8GHz Max, 36MB L3 Cache
GPU	1× NVIDIA GeForce RTX 5090 (32 GB GDDR7)
RAM	192 GB DDR4 ECC DIMM

C.5 Evaluation Metrics Implementation

Expected Calibration Error. In classification settings, a well-calibrated prediction is expected to have a confidence value of 80% and be correct in approximately 80% of the test cases. To calculate expected calibration error (ECE), probabilistic predictions are split into a predetermined number Q of bins B of equal confidence range. The ECE is defined as the average absolute difference between accuracy and confidence over G equally spaced bins (Guo et al., 2017):

$$\mathrm{ECE} := \sum_{g=1}^{G} \frac{|B_g|}{n} \Big| \mathrm{acc}(B_g) - \mathrm{conf}(B_g) \Big|, \tag{A46}$$

where $|B_g|$ is the number of samples in the g-th bin and n is the total number of samples. Following standard practice, we use $G = 10$ bins for evaluation.

Selective Classification. Selective classification evaluates epistemic uncertainty (EU) quality by leveraging the expectation that misclassified samples receive higher EU values than correctly classified ones. Instances with high EU quantification are abstained and deferred to an expert to reduce the risk of misclassification (Algorithm 1). We quantify performance using the accuracy–rejection curve (HÜhn & HÜllermeier, 2009), which plots accuracy on retained samples against the rejection rate. Reliable uncertainty estimates yield a monotonically increasing curve, whereas random rejection yields a flat curve. The area under this curve (AUARC) serves as a scalar summary (Jaeger et al., 2023), with larger values indicating better selective classification performance.

Algorithm 1 selective classification procedure

Input: Test dataset $\mathcal{D}_{\text{test}}$; rejection rate β; epistemic-uncertainty-aware predictor $h(\cdot)$
1. Quantify epistemic predictive uncertainty given each sample, denoted by $u_{\text{EU},n}$
$u_{\text{EU},n} \leftarrow h(x_n)\ \forall n$
2. Sort uncertainty estimates in ascending order
let $\pi = \{1, 2, ..., N_t\}$ with $N_t = |\mathcal{D}_{\text{test}}|$ so that $u_{\text{EU},\pi(1)} \leq u_{\text{EU},\pi(2)} \leq ... \leq u_{\text{EU},\pi(N_t)}$
3. Select top $\lfloor (1-\beta)N_t \rfloor$ certain averaged probabilities
$\tilde{p}_{\pi(1)}, ..., \tilde{p}_{\pi(\lfloor (1-\beta)N_t \rfloor)}$ for class prediction

Balanced Quality Score. To *strike a balance among accuracy, calibration, and selective classification performance*, we introduce the *balanced quality score* (BQS). For each seed, we independently

normalise ACC, $1-\mathrm{ECE}$, and AUARC to $[0,1]$ via min-max scaling across all compared methods within each dataset. BQS is then computed as the average of the three normalised components:

$$\mathrm{BQS} = \frac{1}{3}\Big(\widehat{\mathrm{ACC}} + \widehat{(1-\mathrm{ECE})} + \widehat{\mathrm{AUARC}}\Big), \tag{A47}$$

where each $\widehat{\cdot}$ denotes the min-max normalised value computed across all compared methods for that seed and dataset. We report the mean and standard deviation of BQS across seeds.

D Additional Experiment Results

We report performance comparisons and accuracy–rejection curves under the three annotation imprecision scenarios—annotator disagreement, teacher predictions, and label smoothing—in Tables A4, A5, and A6, and Figures A1, A2, and A3, respectively.

Table A7 further reports results using the teachers' original predictions as the source of annotation imprecision. These results consistently confirm the robustness of POCC in achieving the highest BQS scores across settings.

Table A4: Performance comparison (in %) under annotator-disagreement annotation imprecision across various datasets, averaged over 10 runs.

	MiceBone Dataset				TreeVersity Dataset			
Method	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
EDL	88.5±0.8	9.9±0.8	**97.4±0.3**	68.7±3.6	85.5±0.6	17.1±0.7	95.7±0.3	58.4±1.9
Decali	88.7±1.2	5.8±1.0	93.6±0.9	51.2±6.0	**86.9±0.4**	**2.3±0.3**	95.0±0.4	92.4±2.2
DAPPr	89.0±1.1	13.2±0.9	96.4±0.5	53.6±5.1	86.6±0.5	17.0±0.4	95.5±0.3	61.5±1.5
Ours	**89.1±0.9**	**2.7±0.4**	97.3±0.3	**96.1±4.6**	86.2±0.6	2.5±0.5	**96.7±0.2**	**97.1±1.9**
DE	85.5±0.9	7.5±0.8	94.3±0.5	23.6±5.5	75.7±1.1	6.2±0.9	88.6±0.8	24.4±1.7

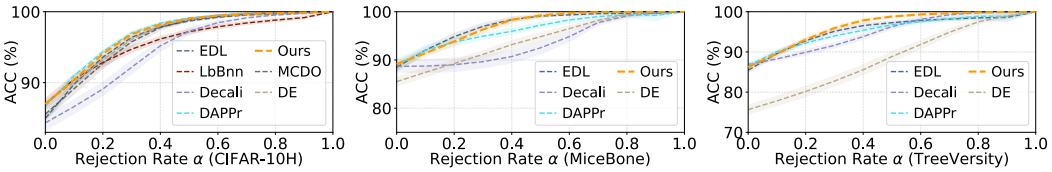


Figure A1: Accuracy rejection curves under annotator-disagreement annotation imprecision.

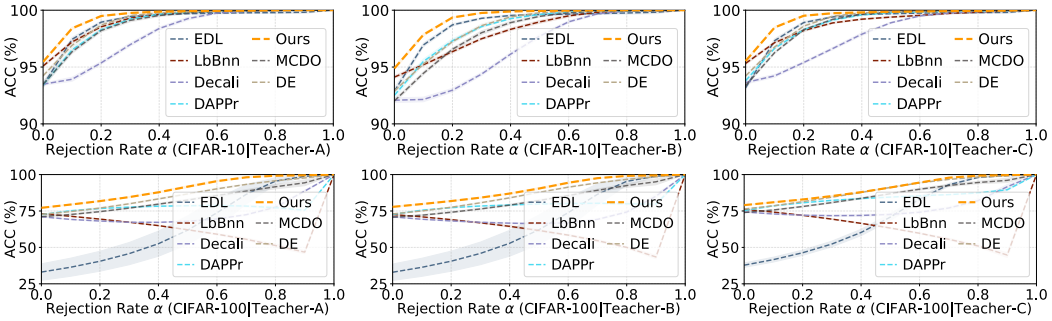


Figure A2: Accuracy rejection curves under teacher-prediction annotation imprecision.

Table A5: Performance comparison (in %) under teacher-prediction annotation imprecision with varying settings, averaged over 10 runs.

Method	CIFAR-10				CIFAR-100			
	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
Teacher: ResNet-20								
EDL	93.4±0.2	9.0±0.2	99.1±0.1	51.6±3.8	33.1±5.7	23.3±3.9	65.8±6.1	30.2±2.9
LbBnn	95.1±0.1	28.1±0.2	99.2±0.0	54.2±3.3	73.1±0.2	63.1±0.2	63.4±0.4	31.9±2.4
Decali	93.5±0.1	8.5±0.1	98.0±0.1	29.7±2.5	71.0±0.2	27.8±0.2	73.3±0.2	64.2±1.9
DAPPr	93.4±0.2	14.6±0.2	99.0±0.0	42.8±3.1	72.5±0.4	52.8±0.3	78.3±0.3	55.1±1.6
Ours	**95.5**±0.1	**3.0**±0.1	**99.5**±0.0	**99.9**±0.3	**77.2**±0.2	**10.6**±0.4	**90.6**±0.2	**100.0**±0.0
MCDO	93.5±0.2	8.5±0.2	99.0±0.1	50.2±4.8	71.1±0.3	28.2±0.3	83.2±0.7	75.5±1.3
DE	94.2±0.1	9.5±0.1	99.2±0.0	65.3±3.0	72.7±0.3	29.9±0.3	86.5±0.3	79.5±0.7
Teacher: ShuffleNetV2-0.5×								
EDL	92.9±0.3	9.1±0.1	98.9±0.1	67.0±3.2	33.0±5.6	23.3±3.8	65.9±6.3	31.9±3.6
LbBnn	94.1±0.1	33.2±0.2	98.3±0.1	43.3±1.8	72.6±0.2	65.4±0.2	62.6±0.4	31.0±2.4
Decali	92.1±0.2	12.1±0.2	96.8±0.1	26.9±1.7	71.0±0.2	27.4±0.2	71.9±0.3	64.0±1.7
DAPPr	92.5±0.2	18.4±0.2	98.6±0.1	47.3±3.8	72.6±0.4	52.3±0.4	79.7±0.2	58.8±1.2
Ours	**94.9**±0.1	**5.4**±0.2	**99.4**±0.0	**100.0**±0.0	**77.8**±0.3	**11.7**±0.3	**90.2**±0.3	**100.0**±0.0
MCDO	92.0±0.2	12.2±0.2	98.3±0.1	45.3±2.0	70.8±0.4	27.7±0.3	83.2±0.6	76.7±1.4
DE	92.9±0.1	13.1±0.1	98.7±0.0	58.5±2.1	73.0±0.2	29.9±0.1	87.2±0.2	81.6±0.9
Teacher: MobileNetV2-x0-5								
EDL	93.2±0.2	9.1±0.2	99.1±0.1	53.1±1.1	37.8±1.5	26.6±1.1	71.4±1.4	32.9±1.3
LbBnn	95.3±0.1	37.4±0.3	99.0±0.0	52.5±2.3	76.1±0.2	68.2±0.3	64.5±0.4	31.0±0.3
Decali	93.6±0.1	10.3±0.1	97.9±0.0	33.7±2.1	74.4±0.3	25.8±0.3	77.4±0.6	70.5±1.0
DAPPr	93.8±0.2	16.1±0.2	99.0±0.0	53.3±2.8	75.3±0.4	51.9±0.4	84.6±0.2	65.2±0.6
Ours	**95.5**±0.1	**4.2**±0.2	**99.5**±0.0	**99.8**±0.4	**79.0**±0.2	**10.9**±0.2	**90.9**±0.2	**100.0**±0.0
MCDO	93.4±0.1	10.2±0.2	99.0±0.0	52.5±2.5	73.9±0.3	25.8±0.2	86.5±0.5	81.7±1.1
DE	94.2±0.1	11.4±0.1	99.1±0.0	66.9±1.4	76.4±0.1	28.4±0.1	90.0±0.1	86.5±0.6

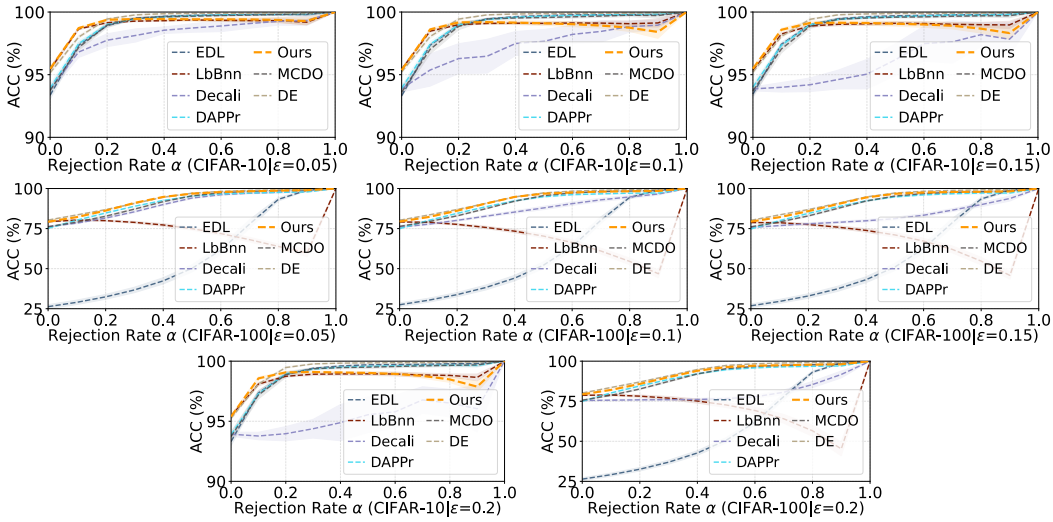


Figure A3: Accuracy rejection curves under label-smoothing annotation imprecision with varying ϵ.

Table A6: Label-smoothing results (%) across ϵ. POCC achieves the highest mean BQS in seven of eight settings; the sole exception is $\epsilon = 0.1$ on CIFAR-10, where it ranks second behind DE (5 members), with the gap attributable to marginal differences in accuracy and AUARC. Statistical significance tests (Table A10, Appendix E) confirm that POCC ranks first statistically across all settings.

	Method	CIFAR-10				CIFAR-100			
		ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
$\epsilon = 0.05$	EDL	93.4±0.3	9.1±0.2	99.1±0.1	48.6±4.3	26.3±1.2	18.7±1.0	58.5±1.6	24.3±0.6
	LbBnn	95.4±0.2	29.9±0.3	99.1±0.1	53.6±2.6	79.3±0.3	61.6±0.5	74.5±0.9	47.7±1.1
	Decali	93.9±0.3	3.6±0.1	98.3±0.2	39.0±5.2	76.2±0.3	3.5±0.4	90.7±0.3	94.0±0.7
	DAPPr	93.9±0.3	8.9±0.2	99.1±0.1	58.5±6.4	75.1±0.4	35.3±0.4	91.9±0.2	76.5±0.4
	Ours	**95.5±0.1**	**1.5±0.1**	99.2±0.1	**90.6±4.7**	79.3±0.2	**2.7±0.2**	93.4±0.1	**99.0±0.3**
	MCDO	93.7±0.2	3.5±0.1	99.1±0.1	60.2±6.3	76.1±0.3	3.8±0.4	91.8±0.2	94.9±0.6
	DE	95.2±0.1	5.5±0.1	**99.4±0.0**	89.7±2.3	**80.3±0.2**	9.1±0.2	**93.8±0.1**	96.4±0.2
$\epsilon = 0.1$	EDL	93.3±0.1	9.2±0.3	99.1±0.0	56.4±2.8	27.5±1.1	19.6±0.9	60.0±1.4	24.6±0.4
	LbBnn	**95.4±0.1**	45.2±0.1	98.9±0.1	56.4±3.6	79.1±0.2	70.0±0.3	69.4±1.2	41.9±1.5
	Decali	93.9±0.3	7.3±0.1	97.5±0.5	39.4±3.6	75.9±0.5	5.0±0.3	87.7±0.7	89.4±0.7
	DAPPr	93.9±0.2	11.8±0.2	99.1±0.1	63.9±4.8	75.0±0.4	36.8±0.4	91.5±0.3	77.1±0.4
	Ours	**95.4±0.1**	**3.7±0.2**	98.8±0.2	87.2±4.4	79.4±0.2	**1.5±0.2**	93.3±0.1	**98.8±0.2**
	MCDO	93.6±0.3	7.8±0.2	99.0±0.1	60.5±5.3	76.4±0.3	7.3±0.3	91.8±0.2	92.7±0.4
	DE	95.2±0.1	10.0±0.1	**99.4±0.0**	**91.2±2.1**	**80.2±0.2**	12.8±0.2	**93.9±0.1**	94.5±0.1
$\epsilon = 0.15$	EDL	93.5±0.3	9.2±0.3	99.1±0.0	61.7±1.7	26.8±1.2	19.1±0.9	59.1±1.5	25.3±0.4
	LbBnn	95.3±0.1	54.5±0.1	98.8±0.1	57.8±2.8	78.8±0.3	72.7±0.3	69.6±1.3	42.5±1.4
	Decali	93.9±0.2	11.6±0.3	96.2±0.7	36.3±3.8	75.7±0.5	8.9±0.3	83.6±0.4	84.1±0.8
	DAPPr	94.0±0.3	15.4±0.2	99.1±0.1	65.7±4.2	74.8±0.5	38.6±0.4	91.5±0.3	77.1±0.4
	Ours	**95.5±0.1**	**6.0±0.2**	98.8±0.2	**92.6±2.8**	79.4±0.1	**2.0±0.2**	93.1±0.1	**98.7±0.2**
	MCDO	93.7±0.4	12.2±0.2	99.0±0.1	62.8±4.6	76.1±0.2	10.9±0.2	91.6±0.2	91.1±0.3
	DE	95.2±0.1	14.3±0.1	**99.4±0.0**	89.6±2.4	**80.2±0.2**	17.1±0.3	**93.9±0.1**	92.9±0.2
$\epsilon = 0.2$	EDL	93.3±0.4	9.2±0.1	99.1±0.1	63.8±2.3	26.4±0.9	18.8±0.7	58.7±1.1	25.8±0.4
	LbBnn	**95.4±0.1**	61.0±0.1	98.7±0.1	58.8±2.0	78.7±0.3	74.2±0.3	70.6±1.7	43.7±1.4
	Decali	93.9±0.2	16.0±0.2	95.5±0.7	37.8±3.0	75.4±0.3	13.5±0.3	80.3±0.7	79.2±1.0
	DAPPr	93.9±0.2	18.9±0.2	99.0±0.1	65.3±2.5	75.0±0.4	40.8±0.3	91.4±0.3	76.6±0.5
	Ours	**95.4±0.1**	**7.6±0.1**	98.7±0.1	**91.9±2.5**	79.3±0.3	**2.7±0.3**	92.8±0.2	**98.4±0.3**
	MCDO	93.7±0.4	16.5±0.2	99.0±0.1	64.0±6.8	75.6±0.3	14.6±0.3	91.4±0.2	89.3±0.4
	DE	95.2±0.1	18.5±0.1	**99.4±0.0**	89.3±2.1	**80.1±0.2**	21.3±0.3	**93.9±0.1**	91.3±0.1

Table A7: Performance comparison (in %) using the teachers' original predictions ($T = 1.0$) as the source of annotation imprecision, averaged over 10 runs. These results consistently confirm the robustness of POCC in achieving the highest BQS scores across settings.

Method	CIFAR-10				CIFAR-100			
	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	ACC ↑	ECE ↓	AUARC ↑	BQS ↑
Teacher: ResNet-20								
EDL	93.3±0.2	9.1±0.3	99.1±0.1	1.2±2.4	33.7±5.0	23.7±3.4	66.5±5.5	14.3±3.1
LbBnn	95.4±0.1	1.5±0.1	99.5±0.0	91.4±2.7	77.3±0.3	19.5±0.5	90.3±0.2	81.3±0.7
Decali	93.8±0.2	1.5±0.1	99.2±0.0	42.2±8.0	74.3±0.4	**2.4±0.3**	89.1±0.1	92.9±1.2
DAPPr	94.0±0.2	7.0±0.1	99.2±0.1	27.2±7.4	75.4±0.3	39.7±0.3	90.1±0.3	61.6±1.2
Ours	**95.5±0.1**	1.8±0.1	**99.5±0.0**	**94.5±2.1**	**77.9±0.2**	3.9±0.4	**92.4±0.1**	**98.6±0.4**
MCDO	93.8±0.1	1.1±0.1	99.2±0.0	48.2±7.6	73.8±0.4	2.9±0.3	90.3±0.2	93.5±1.2
DE	94.8±0.0	**0.7±0.1**	99.4±0.0	77.9±3.9	76.2±0.2	5.9±0.2	91.7±0.1	94.6±0.7
Teacher: ShuffleNetV2-0.5×								
EDL	92.7±0.1	9.1±0.2	99.0±0.0	3.0±2.1	32.7±5.9	23.1±4.2	65.6±6.4	14.1±3.8
LbBnn	95.0±0.2	3.1±0.1	99.3±0.0	76.2±3.1	77.7±0.2	21.7±0.4	90.8±0.1	79.0±0.9
Decali	93.2±0.2	**0.6±0.1**	98.9±0.0	40.9±3.1	74.8±0.3	**2.1±0.3**	90.4±0.1	94.0±1.2
DAPPr	93.6±0.2	8.9±0.2	99.2±0.0	28.4±5.4	75.6±0.3	38.5±0.2	91.0±0.1	62.3±0.8
Ours	**95.0±0.1**	0.8±0.1	**99.5±0.0**	**98.4±1.5**	**78.4±0.1**	3.8±0.2	**92.8±0.1**	**98.3±0.4**
MCDO	93.2±0.3	0.9±0.1	99.0±0.1	48.0±6.1	74.2±0.4	2.3±0.4	90.8±0.3	94.0±1.6
DE	94.0±0.1	1.9±0.1	99.2±0.0	65.5±2.3	77.1±0.2	5.9±0.2	92.3±0.1	94.8±0.5
Teacher: MobileNetV2-x0-5								
EDL	93.3±0.2	9.2±0.2	99.1±0.1	2.7±3.4	38.3±1.6	27.0±1.1	71.7±1.4	9.2±1.1
LbBnn	**95.5±0.2**	1.6±0.1	99.5±0.0	91.0±3.2	**79.2±0.3**	20.9±0.6	91.9±0.1	79.2±0.5
Decali	93.8±0.2	1.4±0.1	99.1±0.0	42.8±8.4	76.2±0.2	**2.3±0.2**	91.7±0.1	94.4±0.3
DAPPr	93.9±0.2	7.2±0.1	99.2±0.0	27.7±5.7	76.1±0.3	36.6±0.3	91.8±0.2	61.2±0.6
Ours	95.5±0.2	1.7±0.1	**99.5±0.0**	**94.6±3.6**	79.1±0.3	3.6±0.3	**93.7±0.1**	**98.4±0.5**
MCDO	93.6±0.2	1.0±0.2	99.2±0.1	47.7±6.5	75.4±0.4	2.4±0.2	91.6±0.3	93.5±0.9
DE	94.8±0.1	**0.8±0.1**	99.4±0.0	77.8±2.4	78.9±0.3	5.2±0.2	93.4±0.1	96.4±0.3

E Evaluation on Statistical Significance Tests

Setup. In this section, we conduct pairwise one-sided Wilcoxon signed-rank tests at the 5% significance level across three evaluation criteria: accuracy, $1 - \text{ECE}$, and AUARC. For each ordered pair of methods (m_i, m_j), $i \neq j$, the null hypothesis H_0 assumes no systematic performance difference between m_i and m_j, while the alternative hypothesis H_1 posits that m_i stochastically outperforms m_j. Tests are performed over 10 independent runs, and results are deemed statistically significant when $p < 0.05$, in which case m_i is considered to outperform m_j.

To provide an interpretable summary of the statistical comparisons, we construct a ranking scheme as follows. For each method and each evaluation criterion, we count the number of significant wins and losses across all pairwise comparisons: a significant win contributes $+1$, a significant loss contributes -1, and a non-significant outcome contributes 0. The net score for method m is defined as $\text{wins}(m) - \text{losses}(m)$, reflecting its relative dominance. To obtain a holistic measure of balanced performance, we min-max normalise the net scores of the three criteria and take their average, denoted as *normalised balanced quality score for statistical test* (BQS-ST), reflecting equal weighting across accuracy, calibration, and EU quantification.

Results. Consistent with the main evaluation settings in Section 5, we evaluate our POCC method across three annotation imprecision settings: annotator disagreement, teacher prediction, and label smoothing. Tables A8, A9, and A10 report rankings derived from pairwise statistical comparisons, demonstrating that *POCC consistently achieves the best trade-off performance*—as measured by the highest BQS-ST scores—across prediction accuracy, calibration, and EU quantification. The detailed statistical significance analyses for the three evaluation scenarios are presented in Figures A4, A5, and A6, respectively.

Table A8: Performance comparison (in %, ↑) under annotator disagreements.

Method	CIFAR-10H				MiceBone				TreeVersity			
	ACC	1-ECE	AUARC	BQS-ST	ACC	1-ECE	AUARC	BQS-ST	ACC	1-ECE	AUARC	BQS-ST
EDL	−5.0	−6.0	−1.0	15.2	−1.0	−2.0	**3.0**	58.3	−2.0	−3.0	1.0	30.4
LbBnn	3.0	−2.0	−4.0	47.8	–	–	–	–	–	–	–	–
Decali	−5.0	**5.0**	−6.0	33.3	1.0	2.0	−4.0	52.8	**3.0**	**3.0**	−2.0	75.0
DAPPr	1.0	−4.0	**5.0**	61.6	**2.0**	−4.0	0.0	52.4	**3.0**	−3.0	1.0	54.2
Ours	**4.0**	2.0	**5.0**	**90.9**	**2.0**	**4.0**	**3.0**	**100.0**	0.0	**3.0**	**4.0**	**85.7**
MCDO	−2.0	**5.0**	−1.0	59.6	–	–	–	–	–	–	–	–
DE	**4.0**	0.0	2.0	75.8	−4.0	0.0	−2.0	26.2	−4.0	0.0	−4.0	16.7

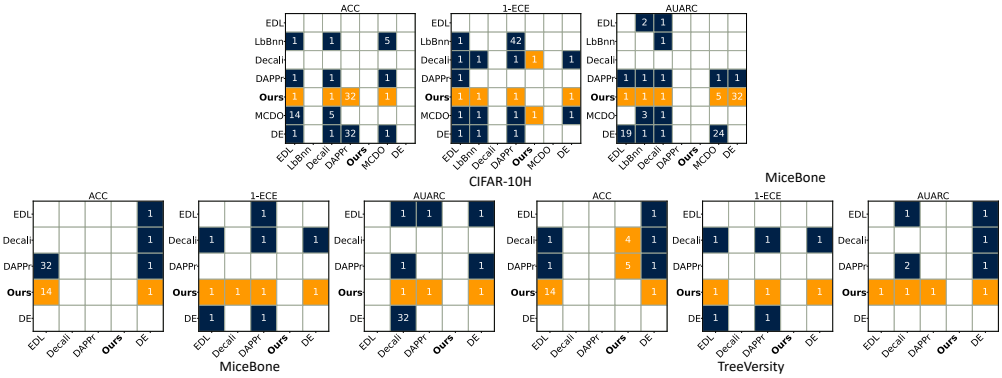


Figure A4: Pairwise one-sided Wilcoxon signed-rank test results on various datasets under annotator disagreement. Cell (i, j) indicates whether row method i significantly outperforms column method j; coloured cells show the scaled p-value ($\times 10^3$). Orange (■) mark comparisons involving our method; dark navy (■) mark all other significant pairs.

Table A9: Performance comparison (in %, ↑) under teacher-prediction annotation imprecision.

	Method	CIFAR-10				CIFAR-100			
		ACC	1-ECE	AUARC	BQS-ST	ACC	1-ECE	AUARC	BQS-ST
ResNet-20	EDL	−3.0	0.0	0.0	33.3	−6.0	4.0	−5.0	27.8
	LbBnn	4.0	−6.0	4.0	53.7	4.0	−6.0	−5.0	27.8
	Decali	−3.0	3.0	−6.0	25.0	−3.0	2.0	−2.0	39.6
	DAPPr	−3.0	−4.0	−3.0	13.9	1.0	−4.0	0.0	40.2
	Ours	**6.0**	**6.0**	**6.0**	**100.0**	**6.0**	**6.0**	**6.0**	**100.0**
	MCDO	−3.0	3.0	−3.0	33.3	−3.0	0.0	2.0	46.2
	DE	2.0	−2.0	2.0	51.9	1.0	−2.0	4.0	57.8
ShuffleNetV2-0.5x	EDL	1.0	4.0	4.0	73.7	−6.0	4.0	−5.0	27.8
	LbBnn	4.0	−6.0	−4.0	32.8	1.0	−6.0	−5.0	19.4
	Decali	−5.0	1.0	−6.0	19.4	−3.0	2.0	−2.0	39.6
	DAPPr	−2.0	−4.0	1.0	34.1	1.0	−4.0	0.0	40.2
	Ours	**6.0**	**6.0**	**6.0**	**100.0**	**6.0**	**6.0**	**6.0**	**100.0**
	MCDO	−5.0	1.0	−2.0	30.6	−3.0	0.0	2.0	46.2
	DE	1.0	−2.0	1.0	48.7	4.0	−2.0	4.0	66.2
MobileNetV2-x0-5	EDL	−6.0	4.0	2.0	50.0	−6.0	2.0	−4.0	27.8
	LbBnn	4.0	−6.0	−3.0	36.1	2.0	−6.0	−6.0	22.2
	Decali	−1.0	1.0	−6.0	33.3	−2.0	2.0	−2.0	44.4
	DAPPr	−1.0	−4.0	−1.0	33.3	0.0	−4.0	0.0	38.9
	Ours	**6.0**	**6.0**	**6.0**	**100.0**	**6.0**	**6.0**	**6.0**	**100.0**
	MCDO	−4.0	1.0	−2.0	36.1	−4.0	2.0	2.0	50.0
	DE	2.0	−2.0	4.0	61.1	4.0	−2.0	4.0	66.7



Figure A5: Pairwise one-sided Wilcoxon signed-rank test results on various datasets under teacher-prediction imprecision. Cell (i, j) indicates whether row method i significantly outperforms column method j; coloured cells show the scaled p-value ($\times 10^3$). Orange (■) mark comparisons involving our method; dark navy (■) mark all other significant pairs.

Table A10: Performance comparison (in %, ↑) under label-smoothing annotation imprecision.

	Method	CIFAR-10				CIFAR-100			
		ACC	1-ECE	AUARC	BQS-ST	ACC	1-ECE	AUARC	BQS-ST
$\epsilon = 0.05$	EDL	−6.0	−4.0	−1.0	19.4	−6.0	−2.0	−6.0	11.1
	LbBnn	**5.0**	−6.0	−2.0	44.4	3.0	−6.0	−4.0	30.6
	Decali	−1.0	3.0	−6.0	40.2	−1.0	3.0	−2.0	50.0
	DAPPr	−2.0	−2.0	2.0	45.5	−4.0	−4.0	1.0	30.6
	Ours	**5.0**	**6.0**	1.0	**86.1**	3.0	**6.0**	4.0	**86.1**
	MCDO	−3.0	3.0	0.0	50.8	−1.0	3.0	1.0	58.3
	DE	2.0	0.0	**6.0**	74.2	**6.0**	0.0	**6.0**	83.3
$\epsilon = 0.1$	EDL	−6.0	0.0	3.0	41.7	−6.0	−2.0	−6.0	11.1
	LbBnn	**5.0**	−6.0	−3.0	41.7	2.0	−6.0	−4.0	27.8
	Decali	−1.0	4.0	−6.0	42.9	−2.0	4.0	−2.0	50.0
	DAPPr	−1.0	−4.0	3.0	45.7	−4.0	−4.0	0.0	27.8
	Ours	**5.0**	**6.0**	−3.0	**75.0**	4.0	**6.0**	4.0	**88.9**
	MCDO	−4.0	2.0	0.0	44.9	0.0	2.0	2.0	61.1
	DE	2.0	−2.0	**6.0**	68.7	**6.0**	0.0	**6.0**	83.3
$\epsilon = 0.15$	EDL	−6.0	4.0	3.0	52.8	−6.0	−2.0	−6.0	11.1
	LbBnn	4.0	−6.0	−3.0	36.1	2.0	−6.0	−4.0	27.8
	Decali	−2.0	2.0	−6.0	33.3	−1.0	4.0	−2.0	52.8
	DAPPr	−2.0	−4.0	3.0	41.7	−4.0	−4.0	1.0	30.6
	Ours	**6.0**	**6.0**	−3.0	**75.0**	4.0	**6.0**	4.0	**88.9**
	MCDO	−2.0	0.0	0.0	44.4	−1.0	2.0	1.0	55.6
	DE	2.0	−2.0	**6.0**	66.7	**6.0**	0.0	**6.0**	83.3
$\epsilon = 0.2$	EDL	−6.0	4.0	4.0	55.6	−6.0	0.0	−6.0	16.7
	LbBnn	**5.0**	−6.0	−3.0	41.7	2.0	−6.0	−4.0	27.8
	Decali	−2.0	2.0	−6.0	34.3	−1.0	4.0	−2.0	52.8
	DAPPr	−2.0	−4.0	1.0	37.1	−4.0	−4.0	1.0	30.6
	Ours	**5.0**	**6.0**	−3.0	**75.0**	4.0	**6.0**	4.0	**88.9**
	MCDO	−2.0	0.0	1.0	48.2	−1.0	2.0	1.0	55.6
	DE	2.0	−2.0	**6.0**	68.7	**6.0**	−2.0	**6.0**	77.8

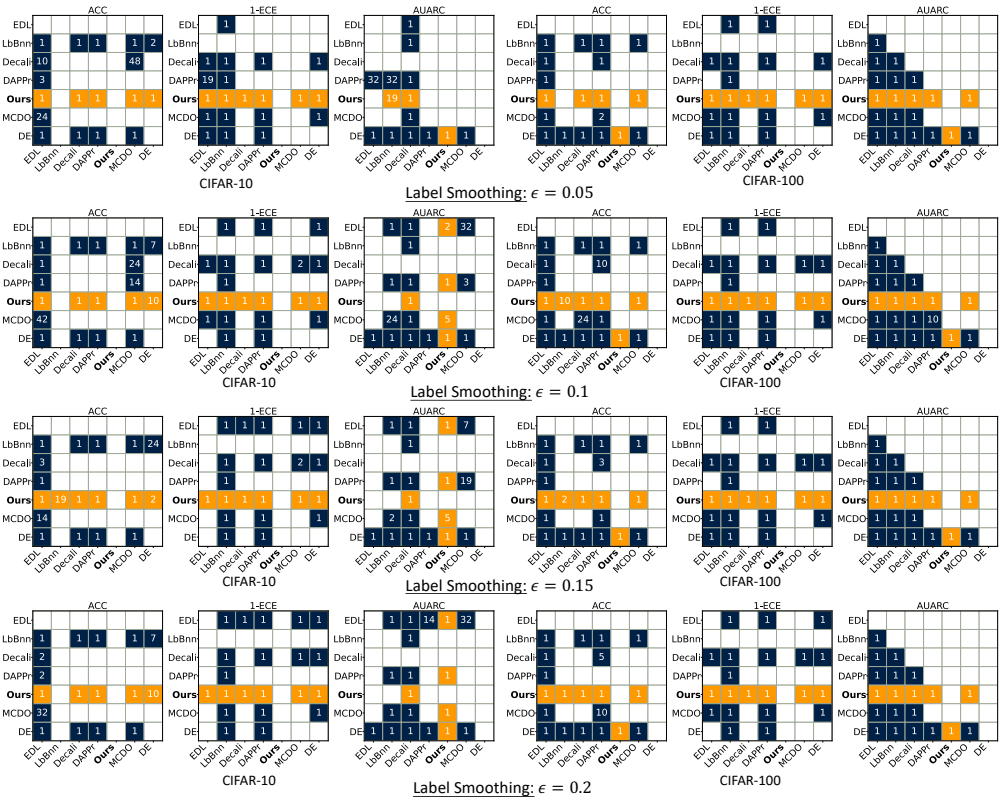


Figure A6: Pairwise one-sided Wilcoxon signed-rank test results on various datasets under label-smoothing imprecision. Cell (i, j) indicates whether row method i significantly outperforms column method j; coloured cells show the scaled p-value ($\times 10^3$). Orange (■) mark comparisons involving our method; dark navy (■) mark all other significant pairs.

F Comparison with Baselines Trained on Collapsed Labels

Setup. In this section, we further compare POCC against baseline methods trained on collapsed hard labels derived from imprecise annotations, covering the annotator-disagreement setting. Specifically, given a soft label q_L, each baseline is trained using the one-hot label e_j, where $j = \arg\max_{k \in [K]} q_L$, i.e., majority voting. All other training configurations follow those of the main evaluation in Section 5, and all results are averaged over 10 independent runs with seeds 1–10.

We exclude the collapsed-label baseline from the knowledge distillation setting, as discarding the teacher's soft predictions contradicts standard practice in knowledge distillation (Hinton et al., 2015; Phuong & Lampert, 2019; Zhou et al., 2021). We exclude the experiment from the label smoothing setting as well, since collapsing a smoothed label simply recovers the original one-hot annotation, containing no annotation imprecision by construction and therefore falling outside the scope of our problem formulation.

Results. Table A11 reports ACC, ECE, AUARC, BQS, and the normalised balanced quality score under statistical significance testing (BQS-ST); see Appendix E for setup details. Across both settings, POCC achieves the highest BQS and BQS-ST values, demonstrating that it consistently yields the best trade-off among prediction accuracy, calibration, and EU quantification compared to baselines trained on collapsed hard labels.

Table A11: Performance comparison (in %) under annotator disagreement, where baseline methods are trained on collapsed hard labels derived from imprecise annotations.

Method	ACC ↑	ECE ↓	AUARC ↑	BQS ↑	BQS-ST ↑
CIFAR-10H					
EDL	84.8±0.2	9.5±0.5	96.6±0.1	41.4±9.2	33.3
LbBnn	85.9±0.5	6.9±0.4	96.2±0.3	53.5±8.9	48.5
Decali	82.7±0.9	9.5±0.5	95.5±0.4	3.4±5.1	0.0
DAPPr	85.5±0.4	6.2±0.4	96.6±0.2	59.5±7.5	62.9
Ours	**87.0±0.4**	2.9±0.3	**97.1±0.2**	**95.7±1.1**	**93.9**
MCDO	83.8±0.8	7.6±0.5	96.0±0.3	28.6±17.8	23.0
DE	85.9±0.4	**1.9±0.2**	96.7±0.2	84.1±10.9	77.8
MiceBone					
EDL	88.5±0.9	9.9±1.0	**97.4±0.3**	58.3±5.4	50.0
Decali	88.5±0.8	4.9±0.7	96.8±0.4	75.0±7.1	52.4
DAPPr	88.5±1.2	8.8±0.6	95.1±0.9	36.0±10.4	26.2
Ours	**89.1±0.9**	**2.7±0.4**	97.3±0.3	**96.2±5.8**	**100.0**
DE	85.6±1.0	2.9±0.7	95.3±0.4	37.3±7.3	33.3
TreeVersity					
EDL	85.4±0.6	17.1±0.7	95.7±0.3	59.5±1.9	39.3
Decali	85.9±0.3	8.5±0.4	95.3±0.3	78.4±2.0	61.9
DAPPr	85.7±0.8	13.2±0.5	94.4±0.5	63.6±1.8	35.7
Ours	**86.2±0.6**	**2.5±0.5**	**96.7±0.2**	**99.4±0.9**	**100.0**
DE	75.1±1.2	3.2±0.6	88.8±0.7	31.7±1.6	25.0

G Evaluation with Normalised AUARC

To assess the quality of EU quantification relative to each model's base classification accuracy in selective classification, we report normalised AUARC values, where accuracy gains are measured relative to the test accuracy without rejection (Wang et al., 2026c). Specifically, let $A(\alpha)$ denote the accuracy at rejection rate α; we define the relative accuracy improvement as $\frac{A(\alpha)-A(0)}{1-A(0)}$, where $A(0)$ is the initial model accuracy without rejection. Under this setting, we compute the balanced quality score for statistical testing (BQS-ST) (see Appendix E for setup details) from ACC, ECE, and normalised AUARC, and report the results in Figures A7, A9, and A8 across the three annotation imprecision scenarios. The results show that our method consistently achieves the best trade-off among accuracy, calibration, and EU quantification.

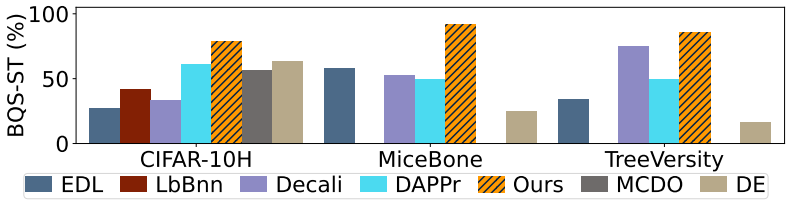


Figure A7: BQS-ST comparison under annotator-disagreement annotation imprecision across various datasets, considering normalised AUARC.

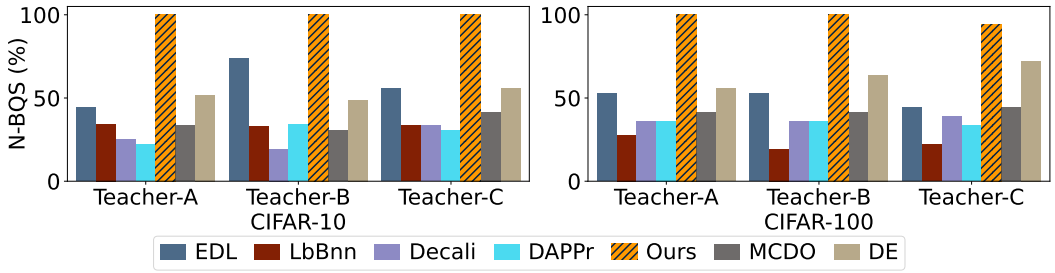


Figure A8: BQS-ST comparison under teacher-prediction annotation imprecision across various settings, considering normalised AUARC. Teacher-A: ResNet-20; Teacher-B: ShuffleNetV2-0.5x; Teacher-C: MobileNetV2-x0-5.

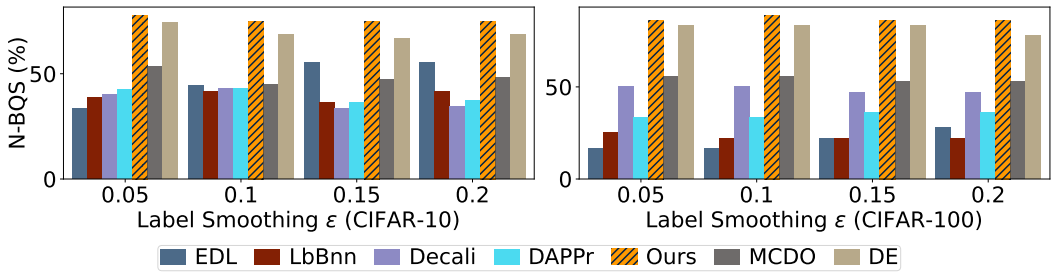


Figure A9: BQS-ST comparison under label-smoothing annotation imprecision with varying ϵ, considering normalised AUARC.

H Evaluation with EU–Prediction Error Correlation

Setup. In this section, we apply an additional evaluation perspective—the alignment between epistemic uncertainty (EU) and prediction error—in place of AUARC in the selective classification task. Specifically, we measure this alignment via the Spearman rank correlation. Let $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the test set. For each sample i, let $u_i \in \mathbb{R}_{\geq 0}$ denote the predicted epistemic uncertainty and $\ell_i = -\log \hat{p}(y_i \mid x_i)$ the cross-entropy loss. The Spearman rank correlation, denoted Sp(EU,CE), is computed as

$$\text{Sp(EU,CE)} = \rho(r(\mathbf{u}),\, r(\boldsymbol{\ell})) = \frac{\text{Cov}(r(\mathbf{u}),\, r(\boldsymbol{\ell}))}{\sigma_{r(\mathbf{u})}\, \sigma_{r(\boldsymbol{\ell})}}, \tag{A48}$$

where $r(\cdot)$ denotes the rank operator, $\mathbf{u} = [u_1, \ldots, u_N]^\top$, and $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_N]^\top$. A higher value indicates stronger rank association with predictive loss: samples assigned higher epistemic uncertainty tend to incur greater prediction error. We adopt rank correlation rather than Pearson correlation because epistemic uncertainty quantities across different methods operate on incomparable scales; ranking removes this dependence while preserving ordinal structure.

A high value of Sp(EU,CE) does not, however, necessarily indicate a well-performing uncertainty estimator: when a model achieves very low accuracy, the cross-entropy loss ℓ_i varies substantially across samples, and a spurious rank alignment with u_i can inflate Sp(EU,CE) even if the uncertainty estimates carry little practical value. This shortcoming highlights the necessity of a balanced evaluation across predictive accuracy, calibration, and Sp(EU,CE). Accordingly, we compute the balanced quality score (BQS) and the score for statistical testing (BQS-ST) from ACC, ECE, and Sp(EU,CE) (see Appendix E for setup details).

Results. Tables A12, A13, and A14 across the three annotation imprecision scenarios. The results show that our method consistently achieves the best trade-off among accuracy, calibration, and EU quantification. In contrast, none of the baselines exhibits a consistent pattern that achieves stable secondary or competitive balance scores across settings. This stems from its explicit treatment of annotation imprecision: baselines trained under standard supervised classification collapse the label into a single distribution, so their EU quantification mainly reflects model-internal uncertainty (e.g., initialisation or approximate inference) rather than supervision ambiguity. In contrast, POCC trains on credal labels and jointly optimises pessimistic and optimistic heads, so the resulting predictive credal set directly encodes annotation ambiguity. This enables simultaneous improvements in accuracy, calibration, and EU quality rather than (significantly) trading one metric against another.

Table A12: Annotator-disagreement results (%). Best in bold; second best underlined.

	Method	ACC ↑	ECE ↓	Sp(EU,CE) ↑	BQS ↑	BQS-ST ↑
CIFAR-10H	EDL	85.0±0.4	9.5±0.3	**95.3±0.2**	40.8±6.4	33.3
	LbBnn	87.0±0.8	6.6±0.3	88.8±0.7	51.7±4.4	47.8
	Decali	84.3±1.0	1.0±0.2	85.6±0.9	35.4±4.5	33.3
	DAPPr	86.6±0.3	6.9±0.2	**95.3±0.2**	68.5±5.6	61.6
	Ours	**87.0±0.4**	2.9±0.3	94.4±0.5	**84.9±5.7**	**81.8**
	MCDO	85.6±0.7	**0.9±0.1**	90.4±0.4	63.8±8.7	56.6
	DE	87.0±0.3	4.3±0.2	92.0±0.2	71.0±4.5	69.7
MiceBone	EDL	88.5±0.8	9.9±0.8	**96.1±0.5**	69.4±3.1	58.3
	Decali	88.7±1.2	5.8±1.0	72.4±0.9	50.2±5.1	52.8
	DAPPr	89.0±1.1	13.2±0.9	87.1±1.3	49.9±4.2	58.3
	Ours	**89.1±0.9**	**2.7±0.4**	82.2±1.8	**77.7±4.2**	**83.3**
	DE	85.5±0.9	7.5±0.8	75.9±2.1	22.6±4.4	25.0
TreeVersity	EDL	85.5±0.6	17.1±0.7	**92.8±0.6**	62.6±1.6	42.9
	Decali	**86.9±0.4**	**2.3±0.3**	81.7±1.7	83.4±2.4	75.0
	DAPPr	86.6±0.5	17.0±0.4	85.8±1.1	56.3±1.5	50.0
	Ours	86.2±0.6	2.5±0.5	90.8±0.5	**94.2±1.8**	**77.4**
	DE	75.7±1.1	6.2±0.9	69.6±1.7	24.4±1.7	16.7

Table A13: Teacher-prediction results (%). Best in bold; second best underlined.

Method	CIFAR-10					CIFAR-100				
	ACC	ECE	Sp(EU,CE)	BQS	BQS-ST	ACC	ECE	Sp(EU,CE)	BQS	BQS-ST
Teacher: ResNet-20										
EDL	93.4±0.2	9.0±0.2	**98.9±0.1**	60.7±2.9	50.0	33.1±5.7	23.3±3.9	62.7±5.6	56.6±1.0	50.0
LbBnn	95.1±0.1	28.1±0.2	96.0±0.1	51.3±2.9	44.1	73.1±0.2	63.1±0.2	−43.4±1.4	30.2±0.3	27.8
Decali	93.5±0.1	8.5±0.1	88.7±0.4	29.7±2.5	25.0	71.0±0.2	27.8±0.2	6.1±0.4	65.7±0.6	36.1
DAPPr	93.4±0.2	14.6±0.2	97.6±0.1	49.8±2.4	29.8	72.5±0.4	52.8±0.3	42.8±0.7	61.7±0.5	36.1
Ours	**95.5±0.1**	**3.0±0.1**	**98.9±0.1**	**99.7±0.6**	**100.0**	**77.2±0.2**	**10.6±0.4**	63.0±1.4	**98.1±0.4**	**91.7**
MCDO	93.5±0.2	8.5±0.2	90.3±0.2	34.5±3.3	31.1	71.1±0.3	28.2±0.3	58.3±1.6	80.9±0.8	44.4
DE	94.2±0.1	9.5±0.1	91.8±0.1	49.4±2.4	41.8	72.7±0.3	29.9±0.3	**69.3±0.6**	84.2±0.4	63.9
Teacher: ShuffleNetV2-0.5×										
EDL	92.9±0.3	9.1±0.1	**98.7±0.1**	73.0±2.5	79.3	33.0±5.6	23.3±3.8	63.1±6.0	57.3±0.8	52.8
LbBnn	94.1±0.1	33.2±0.2	89.8±0.2	46.3±1.4	38.4	72.6±0.2	65.4±0.2	−51.1±1.0	29.3±0.6	19.4
Decali	92.1±0.2	12.1±0.2	73.4±0.3	26.9±1.7	19.4	71.0±0.2	27.4±0.2	−1.1±0.6	65.4±0.6	36.1
DAPPr	92.5±0.2	18.4±0.2	95.7±0.1	53.6±3.1	36.9	72.6±0.4	52.3±0.4	48.6±0.8	64.8±0.6	36.1
Ours	**94.9±0.1**	**5.4±0.2**	98.4±0.1	**99.6±0.2**	**94.4**	**77.8±0.3**	**11.7±0.3**	58.2±1.3	**96.5±0.3**	**88.9**
MCDO	92.0±0.2	12.2±0.2	88.9±0.2	46.6±1.2	25.0	70.8±0.4	27.7±0.3	56.9±1.2	80.9±0.8	44.4
DE	92.9±0.1	13.1±0.1	90.7±0.2	57.6±1.5	46.0	73.0±0.2	29.9±0.1	**71.0±0.5**	85.1±0.5	72.2
Teacher: MobileNetV2-x0-5										
EDL	93.2±0.2	9.1±0.2	**98.9±0.1**	61.8±0.3	61.1	37.8±1.5	26.6±1.1	68.6±1.1	56.0±0.4	50.0
LbBnn	95.3±0.1	37.4±0.3	89.7±0.4	46.5±1.9	44.4	76.1±0.2	68.2±0.3	−62.3±0.8	31.0±0.3	22.2
Decali	93.6±0.1	10.3±0.1	80.9±0.2	33.7±2.1	33.3	74.4±0.3	25.8±0.3	13.5±1.3	72.6±0.5	38.9
DAPPr	93.8±0.2	16.1±0.2	97.1±0.1	60.1±2.4	41.7	75.3±0.4	51.9±0.4	63.0±0.8	70.2±0.3	41.7
Ours	**95.5±0.1**	**4.2±0.2**	98.6±0.1	**99.3±0.5**	**94.4**	**79.0±0.2**	**10.9±0.2**	59.4±1.0	**96.2±0.3**	**77.8**
MCDO	93.4±0.1	10.2±0.2	84.9±0.1	37.4±2.0	30.6	73.9±0.3	25.8±0.2	64.0±1.2	84.5±0.5	47.2
DE	94.2±0.1	11.4±0.1	86.1±0.2	50.4±1.6	44.4	76.4±0.1	28.4±0.1	**75.1±0.4**	87.7±0.3	72.2

Table A14: Label-smoothing results (%). Best in bold; second best underlined.

Method	CIFAR-10					CIFAR-100				
	ACC	ECE	Sp(EU,CE)	BQS	BQS-ST	ACC	ECE	Sp(EU,CE)	BQS	BQS-ST
$\epsilon = 0.05$										
EDL	93.4±0.3	9.1±0.2	**99.0±0.1**	58.9±1.9	38.9	26.3±1.2	18.7±1.0	58.6±1.8	47.3±0.6	16.7
LbBnn	95.4±0.2	29.9±0.3	64.8±2.6	40.3±3.8	45.5	79.3±0.3	61.6±0.5	−5.8±4.0	32.7±0.3	25.0
Decali	93.9±0.3	3.6±0.1	52.3±6.5	39.5±5.6	40.2	76.2±0.3	3.5±0.4	76.6±0.9	93.1±0.7	50.0
DAPPr	93.9±0.3	8.9±0.2	98.9±0.1	67.1±4.0	56.6	75.1±0.4	35.3±0.4	**87.6±0.4**	78.4±0.2	44.4
Ours	**95.5±0.1**	**1.5±0.1**	97.5±0.4	**97.9±1.7**	**90.9**	79.3±0.2	**2.7±0.2**	85.7±0.4	**98.7±0.3**	**80.6**
MCDO	93.7±0.2	3.5±0.1	61.3±1.6	43.6±5.2	40.2	76.1±0.3	3.8±0.4	82.4±0.5	95.0±0.6	55.6
DE	95.2±0.1	5.5±0.1	69.8±0.9	68.8±3.9	59.1	**80.3±0.2**	9.1±0.2	86.8±0.3	96.1±0.2	77.8
$\epsilon = 0.1$										
EDL	93.3±0.1	9.2±0.3	**99.0±0.1**	62.6±0.6	50.0	27.5±1.1	19.6±0.9	60.2±1.1	51.5±0.4	22.2
LbBnn	**95.4±0.1**	45.2±0.1	43.6±6.0	38.3±4.6	33.3	79.1±0.2	70.0±0.3	−52.6±4.3	32.6±0.2	22.2
Decali	93.9±0.3	7.3±0.1	37.1±13.5	42.3±5.9	42.9	75.9±0.5	5.0±0.3	57.2±2.9	88.5±0.8	44.4
DAPPr	93.9±0.2	11.8±0.2	98.9±0.1	69.6±3.4	54.0	75.0±0.4	36.8±0.4	**86.9±0.6**	79.5±0.2	44.4
Ours	95.4±0.1	**3.7±0.2**	95.8±0.5	**96.9±1.8**	**90.0**	79.4±0.2	**1.5±0.2**	83.7±0.4	**98.7±0.2**	**83.3**
MCDO	93.6±0.3	7.8±0.2	58.3±1.4	47.6±4.1	38.3	76.4±0.3	7.3±0.3	80.0±0.3	93.1±0.2	55.6
DE	95.2±0.1	10.0±0.1	68.1±1.3	75.6±2.8	52.0	**80.2±0.2**	12.8±0.2	85.5±0.2	94.2±0.2	77.8
$\epsilon = 0.15$										
EDL	93.5±0.3	9.2±0.3	**98.9±0.1**	65.1±1.4	61.1	26.8±1.2	19.1±0.9	59.5±1.3	52.0±0.5	22.2
LbBnn	95.3±0.1	54.5±0.1	32.2±3.2	35.9±3.7	33.3	78.8±0.3	72.7±0.3	−51.7±3.9	32.5±0.3	22.2
Decali	93.9±0.2	11.6±0.3	21.5±11.6	37.1±4.8	33.3	75.7±0.5	8.9±0.3	34.2±2.1	81.3±0.8	47.2
DAPPr	94.0±0.3	15.4±0.2	98.8±0.1	69.3±4.0	44.4	74.8±0.5	38.6±0.4	**86.9±0.7**	79.4±0.2	44.4
Ours	**95.5±0.1**	**6.0±0.2**	93.9±0.5	**97.6±0.5**	**88.9**	79.4±0.1	**2.0±0.2**	81.5±0.6	**98.2±0.2**	**83.3**
MCDO	93.7±0.4	12.2±0.2	55.2±1.4	48.7±3.2	38.9	76.1±0.2	10.9±0.2	77.7±0.4	91.1±0.2	52.8
DE	95.2±0.1	14.3±0.1	66.1±1.6	75.6±2.4	50.0	**80.2±0.2**	17.1±0.3	84.5±0.2	92.3±0.2	77.8
$\epsilon = 0.2$										
EDL	93.3±0.4	9.2±0.1	**98.9±0.1**	66.9±2.0	61.1	26.4±0.9	18.8±0.7	59.2±1.5	52.4±0.3	27.8
LbBnn	**95.4±0.1**	61.0±0.1	26.2±4.6	37.0±4.4	33.3	78.7±0.3	74.2±0.3	−47.4±5.4	32.5±0.2	22.2
Decali	93.9±0.2	16.0±0.2	18.0±18.5	40.3±4.5	34.3	75.4±0.3	13.5±0.3	17.1±2.1	74.8±0.8	47.2
DAPPr	93.9±0.2	18.9±0.2	98.7±0.1	68.8±2.2	44.9	75.0±0.4	40.8±0.3	**86.6±0.6**	79.0±0.2	44.4
Ours	95.4±0.1	**7.6±0.1**	91.6±0.6	**95.6±1.9**	**87.9**	79.3±0.3	**2.7±0.3**	79.2±0.7	**97.6±0.2**	**83.3**
MCDO	93.7±0.4	16.5±0.2	52.9±1.5	49.9±5.6	37.9	75.6±0.3	14.6±0.3	75.4±0.6	88.9±0.4	52.8
DE	95.2±0.1	18.5±0.1	65.1±0.9	76.1±3.2	50.5	**80.1±0.2**	21.3±0.3	84.1±0.3	90.7±0.2	72.2

I Ablations on Epistemic Uncertainty Quantification

I.1 Maximum Mean Imprecision

The exact maximum mean imprecision (MMI) for measuring EU under total variation (Chau et al., 2025a) is defined as

$$\mathrm{MMI}(\underline{P}) := \sup_{A \in 2^{\mathcal{Y}}} \big(\overline{P}(A) - \underline{P}(A)\big), \tag{A49}$$

where $2^{\mathcal{Y}}$ denotes the power set over the output space $\mathcal{Y}$, and the upper and lower probabilities $\overline{P}(A)$ and $\underline{P}(A)$ bound the probability of any event $A \subseteq \mathcal{Y}$. Given our credal set defined by eq. (6), i.e.,

$$C_\theta = \big\{(1-\epsilon)\, p_{+,\theta} + \epsilon\, p_{-,\theta} \mid \epsilon \in [0,1]\big\},$$

the induced probability $p_\epsilon(A)$ of any event A is affine in ϵ, so its extrema over C are attained at $\epsilon \in \{0,1\}$, as follows:

$$\overline{P}(A) = \max\big\{p_{+,\theta}(A),\, p_{-,\theta}(A)\big\}, \qquad \underline{P}(A) = \min\big\{(p_{+,\theta}(A),\, p_{-,\theta}(A)\big\}.$$

Therefore, the MMI in eq. (A49) reduces exactly to the total variation distance between the two predictive distributions:

$$\mathrm{MMI}(\underline{P}) = \sup_{A \in 2^{\mathcal{Y}}} \big|p_{+,\theta}(A) - p_{-,\theta}(A)\big| = \frac{1}{2}\sum\nolimits_{k=1}^{K} \big|p_{+,\theta,k} - p_{-,\theta,k}\big|, \tag{A50}$$

where the last equality is the standard characterisation of the total variation distance in terms of the underlying probability mass functions.

I.2 Shannon Entropy Difference

Since reliable uncertainty quantification depends jointly on the representation and the uncertainty measure (Wang et al., 2026d), in addition to MMI, we also consider the Shannon entropy difference ($\mathrm{H}_{\mathrm{diff}}$) (Abellán et al., 2006) as an alternative, defined as the range of entropy over the credal set:

$$\mathrm{H}_{\mathrm{diff}}(x) = \max_{p \in C_\theta(x)} H(p) - \min_{p \in C_\theta(x)} H(p), \tag{A51}$$

where $H(p) := -\sum_{k=1}^{K} p_k \log_2 p_k$ denotes the Shannon entropy of a predictive distribution. Unlike MMI, however, the maximisation in eq. (A51) does not admit a closed-form solution and requires numerical optimisation, incurring a small additional computational overhead.

Experimental Evaluation. We compare MMI and $\mathrm{H}_{\mathrm{diff}}$ for EU quantification using our POCC method on AUARC in selective classification under three annotation imprecision scenarios: annotator disagreement (Table A15), teacher-model predictions (Table A16), and label smoothing (Table A17), corresponding to the evaluation settings in Section 5. The results show that MMI consistently achieves better EU quantification performance while being computationally more efficient.

Table A15: Comparison of MMI and $\mathrm{H}_{\mathrm{diff}}$ on AUARC (↑) in selective classification under annotator-disagreement annotation imprecision.

Metric	CIFAR-10H	MiceBone	TreeVersity
MMI	$97.1_{\pm 0.2}$	$97.3_{\pm 0.3}$	$96.7_{\pm 0.2}$
H-Diff	$96.6_{\pm 0.2}$	$95.0_{\pm 0.7}$	$95.7_{\pm 0.2}$

Table A16: Comparison of MMI and $\mathrm{H}_{\mathrm{diff}}$ on AUARC (↑) in selective classification under annotation imprecision from teacher model predictions.

	ShuffleNetV2-0.5x		ResNet-20		MobileNetV2-x0-5	
Dataset	MMI	H-Diff	MMI	H-Diff	MMI	H-Diff
CIFAR-10	$99.4_{\pm 0.0}$	$99.1_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.3_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.2_{\pm 0.0}$
CIFAR-100	$90.2_{\pm 0.3}$	$83.0_{\pm 0.7}$	$90.6_{\pm 0.2}$	$84.5_{\pm 0.6}$	$90.9_{\pm 0.2}$	$83.7_{\pm 0.4}$

Table A17: Comparison of MMI and H_{diff} on AUARC (↑) in selective classification under annotation imprecision from label smoothing.

Dataset	Smoothing ϵ	MMI	H-Diff
CIFAR-10	0.05	$99.2_{\pm 0.1}$	$99.1_{\pm 0.1}$
	0.10	$98.8_{\pm 0.2}$	$98.6_{\pm 0.2}$
	0.15	$98.8_{\pm 0.2}$	$98.1_{\pm 0.2}$
	0.20	$98.7_{\pm 0.1}$	$97.3_{\pm 0.2}$
CIFAR-100	0.05	$93.4_{\pm 0.1}$	$92.2_{\pm 0.1}$
	0.10	$93.3_{\pm 0.1}$	$91.5_{\pm 0.2}$
	0.15	$93.1_{\pm 0.1}$	$90.7_{\pm 0.2}$
	0.20	$92.8_{\pm 0.2}$	$89.6_{\pm 0.3}$

Robustness of POCC on EU Quantification Measures. When using H_{diff} as the EU measure, our method consistently achieves the best balanced quality scores (BQSs) across all scenarios (Figures A10, A11), and A12, with one exception: on CIFAR-10 under label smoothing, it ranks second, behind the 5-member Deep Ensemble.

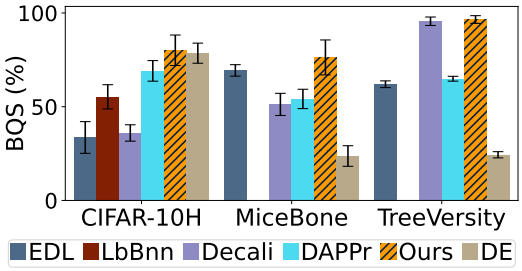


Figure A10: BQS of our method under annotator-disagreement annotation imprecision, using H_{diff} for EU quantification.

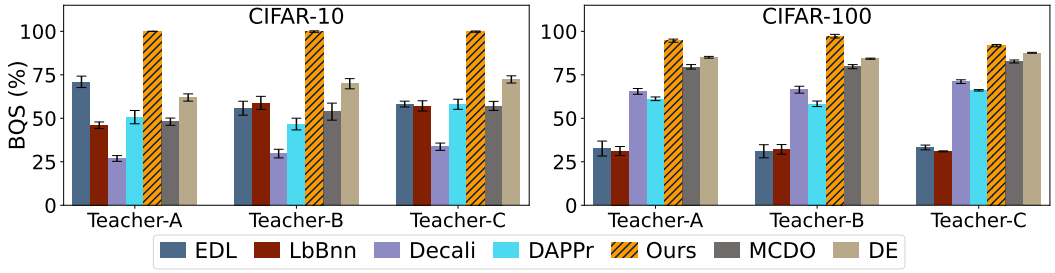


Figure A11: BQS of our method under teacher-prediction annotation imprecision, using H_{diff} for EU quantification. Teacher-A: ResNet-20; Teacher-B: ShuffleNetV2-0.5x; Teacher-C: MobileNetV2-x0-5.

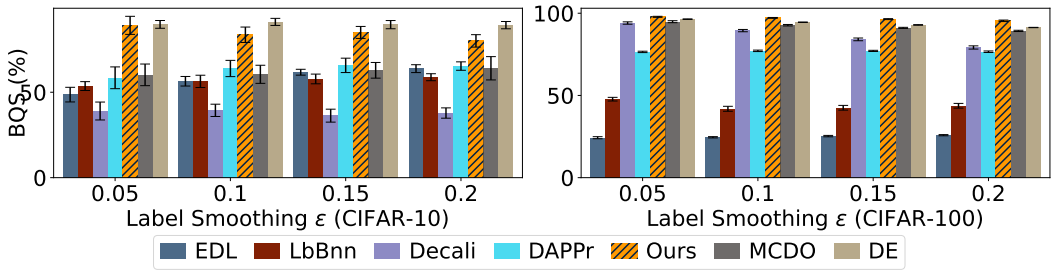


Figure A12: BQS of our method under label-smoothing annotation imprecision with varying ϵ, using H_{diff} for EU quantification.

J On the Exclusion of OOD Detection for EU Evaluation

Below we detail why out-of-distribution (OOD) detection is not an appropriate benchmark for EU assessment in our work.

Conceptual Mismatch. OOD detection asks whether a test input x was drawn from the same marginal distribution P_X as the training data—an *input-space* question. The EU studied in our framework asks something fundamentally different: *given x, how much remains underdetermined about the true conditional label distribution because the supervision itself is imprecise?*—a *label-space* question. These two sources of uncertainty are conceptually orthogonal. A perfectly in-distribution image can receive highly divergent human annotations (high annotation-induced EU, yet assuredly in-distribution), while a clearly OOD input can be unambiguous to any annotator (low annotation-induced EU, yet OOD). Using OOD-detection performance as a proxy for EU therefore conflates two distinct epistemic phenomena that need not correlate.

Methodological Critique. This mismatch is not unique to our setting. Li et al. (2025) argue that supervised classifiers answer the wrong question for OOD detection: high predictive uncertainty and large feature distance are conflated with being OOD, yielding irreducible errors that neither scaling, epistemic uncertainty representations, nor outlier exposure can remove. The critique applies with particular force to POCC, whose EU measures the size of the predictive credal set induced by imprecise annotations. Nothing in its training objective ties this quantity to deviation from P_X.

Absence of a Meaningful Reference Signal. Even setting aside the above, standard OOD benchmarks provide no reference signal for annotation-induced EU. The proxy rests on the assumption that OOD samples *should* exhibit higher EU than in-distribution ones—an assumption that is debatable in conventional settings (Li et al., 2025) and untenable here: all test samples in our experiments are drawn from the same input distribution as the training data. EU differences arise from variation in the informativeness and consistency of supervision, not from variation in input provenance, so an OOD benchmark cannot distinguish between instances that are genuinely annotation-ambiguous and those that are not.

Selective classification, by contrast, offers a task-oriented evaluation with a clear operational interpretation aligned with our framework's objective. A good EU estimate should rank genuinely annotation-ambiguous instances higher, so that abstaining on high-uncertainty predictions improves accuracy on the retained subset. This provides operational evidence that the learned disagreement between pessimistic and optimistic objectives is useful for ranking prediction errors—without claiming that the predictive segment faithfully recovers the true credal set induced by annotation imprecision, or that OOD detection serves as a valid proxy for EU.

Additional Experiments. Although OOD detection is not an appropriate benchmark for EU evaluation in our work (as discussed above), one may still be interested in how our method performs on OOD detection. We therefore conduct the following additional experiments.

We consider models trained under three annotation imprecision settings: (i) CIFAR-10H; (ii) CIFAR-10 with ResNet-20 as the teacher in a knowledge distillation setting; and (iii) CIFAR-10 with label smoothing at $\epsilon = 0.05$. For OOD detection, CIFAR-10 is used as the in-distribution dataset, with several OOD datasets: SVHN (Hendrycks et al., 2021), Places365 (Zhou et al., 2018), FMNIST (Xiao et al., 2017), and ImageNet (Deng et al., 2009). For a fair comparison with multiple-forward-pass baselines, we additionally include an ensemble of 5 POCC models, denoted Our-Ens.

Let M denote the number of independently trained POCC models (via random initialisation, same as DE) with parameters $\theta^{(1)}, \ldots, \theta^{(M)}$. Each model m produces two vertices in the probability simplex via its pessimistic and optimistic heads, $p_{-,\theta^{(m)}}(x), \quad p_{+,\theta^{(m)}}(x) \in \Delta^{K-1}$, which together define a line-segment credal set $C_{\theta^{(m)}}(x)$ as in eq. (6). Aggregating across all M models, we collect the full set of $2M$ vertices,

$$\mathcal{V}(x) \;=\; \bigcup_{m=1}^{M} \left\{ p_{-,\theta^{(m)}}(x),\; p_{+,\theta^{(m)}}(x) \right\} \;\subset\; \Delta^{K-1}. \tag{A52}$$

The ensemble credal set is then defined as the convex hull of $\mathcal{V}(x)$,

$$\mathrm{conv}(\mathcal{V}(x)) = \left\{ \sum_{m=1}^{M} \left(\alpha_m^- \, p_{-,\theta^{(m)}}(x) + \alpha_m^+ \, p_{+,\theta^{(m)}}(x) \right) \middle| \alpha_m^-, \alpha_m^+ \geq 0, \sum_{m=1}^{M} \left(\alpha_m^- + \alpha_m^+ \right) = 1 \right\}, \tag{A53}$$

which is a convex polytope in Δ^{K-1} with at most $2M$ vertices. EU is quantified as the Shannon entropy difference in eq. (A51).

Results. Tables A18, A19, and A20 report the OOD detection results. POCC achieves comparable performance to the single-model baselines (ranking within the top 2), while the POCC ensemble attains the best OOD detection performance overall.

Table A18: AUROC (%) for OOD detection using epistemic uncertainty, with CIFAR-10 (models trained using CIFAR-10H) as the ID dataset, averaged over 10 runs. Methods are grouped into *single-model* approaches (above the line) and *multiple-forward-pass* approaches, including an ensemble of our POCCs (Our-Ens), MCDO, and DE (below the line). Bold and underlined entries denote the best and second-best results within each group, respectively.

	SVHN	Places	FMNIST	ImageNet
EDL	$0.735_{\pm 0.004}$	$0.717_{\pm 0.004}$	$0.721_{\pm 0.007}$	$0.701_{\pm 0.004}$
Decali	$\mathbf{0.854_{\pm 0.011}}$	$\mathbf{0.832_{\pm 0.012}}$	$\mathbf{0.846_{\pm 0.014}}$	$\mathbf{0.807_{\pm 0.011}}$
DAPPr	$0.730_{\pm 0.007}$	$0.707_{\pm 0.006}$	$0.699_{\pm 0.009}$	$0.680_{\pm 0.006}$
Ours	$\underline{0.840_{\pm 0.009}}$	$\underline{0.816_{\pm 0.009}}$	$\underline{0.830_{\pm 0.013}}$	$\underline{0.785_{\pm 0.007}}$
Our-Ens	$\mathbf{1.000_{\pm 0.000}}$	$\mathbf{0.992_{\pm 0.001}}$	$\mathbf{0.997_{\pm 0.001}}$	$\mathbf{0.990_{\pm 0.001}}$
MCDO	$0.957_{\pm 0.009}$	$0.930_{\pm 0.009}$	$0.945_{\pm 0.009}$	$0.904_{\pm 0.010}$
DE	$\underline{0.981_{\pm 0.003}}$	$\underline{0.958_{\pm 0.004}}$	$\underline{0.971_{\pm 0.003}}$	$\underline{0.937_{\pm 0.005}}$

Table A19: AUROC (%) for OOD detection using epistemic uncertainty, with CIFAR-10 (model trained using the ResNet-20 as the teacher model with $T = 2.5$) as the ID dataset, averaged over 10 runs. Methods are grouped into *single-model* approaches (above the line) and *multiple-forward-pass* approaches, including an ensemble of our POCCs (Our-Ens), MCDO, and DE (below the line). Bold and underlined entries denote the best and second-best results within each group, respectively.

	SVHN	Places	FMNIST	ImageNet
EDL	$0.717_{\pm 0.008}$	$0.696_{\pm 0.006}$	$0.693_{\pm 0.010}$	$0.670_{\pm 0.005}$
Decali	$\mathbf{0.750_{\pm 0.003}}$	$\underline{0.713_{\pm 0.003}}$	$\mathbf{0.737_{\pm 0.005}}$	$\mathbf{0.694_{\pm 0.002}}$
DAPPr	$0.609_{\pm 0.004}$	$0.605_{\pm 0.003}$	$0.603_{\pm 0.003}$	$0.601_{\pm 0.003}$
Ours	$\underline{0.743_{\pm 0.003}}$	$\mathbf{0.721_{\pm 0.003}}$	$\underline{0.730_{\pm 0.006}}$	$\underline{0.688_{\pm 0.003}}$
Our-Ens	$\mathbf{1.000_{\pm 0.000}}$	$\mathbf{0.989_{\pm 0.001}}$	$\mathbf{0.997_{\pm 0.000}}$	$\mathbf{0.980_{\pm 0.001}}$
MCDO	$0.944_{\pm 0.001}$	$0.902_{\pm 0.002}$	$0.940_{\pm 0.004}$	$0.852_{\pm 0.002}$
DE	$\underline{0.968_{\pm 0.001}}$	$\underline{0.917_{\pm 0.001}}$	$\underline{0.956_{\pm 0.003}}$	$\underline{0.865_{\pm 0.002}}$

Table A20: AUROC (%) for OOD detection using epistemic uncertainty, with CIFAR-10 (model trained using label smoothing $\varepsilon = 0.05$) as the ID dataset, averaged over 10 runs. Methods are grouped into *single-model* approaches (above the line) and *multiple-forward-pass* approaches, including an ensemble of our POCCs (Our-Ens), MCDO, and DE (below the line). Bold and underlined entries denote the best and second-best results within each group, respectively.

	SVHN	Places	FMNIST	ImageNet
EDL	$0.715_{\pm 0.008}$	$0.696_{\pm 0.006}$	$0.691_{\pm 0.008}$	$0.671_{\pm 0.004}$
Decali	$\underline{0.814_{\pm 0.020}}$	$\underline{0.780_{\pm 0.026}}$	$\underline{0.784_{\pm 0.020}}$	$\underline{0.755_{\pm 0.013}}$
DAPPr	$0.744_{\pm 0.003}$	$0.706_{\pm 0.002}$	$0.720_{\pm 0.007}$	$0.664_{\pm 0.002}$
Ours	$\mathbf{0.872_{\pm 0.009}}$	$\mathbf{0.836_{\pm 0.008}}$	$\mathbf{0.846_{\pm 0.007}}$	$\mathbf{0.798_{\pm 0.010}}$
Our-Ens	$\mathbf{0.999_{\pm 0.000}}$	$\mathbf{0.989_{\pm 0.001}}$	$\mathbf{0.997_{\pm 0.000}}$	$\mathbf{0.976_{\pm 0.001}}$
MCDO	$0.920_{\pm 0.002}$	$0.891_{\pm 0.002}$	$0.906_{\pm 0.005}$	$0.861_{\pm 0.003}$
DE	$\underline{0.947_{\pm 0.001}}$	$\underline{0.911_{\pm 0.001}}$	$\underline{0.933_{\pm 0.003}}$	$\underline{0.884_{\pm 0.001}}$