Structured Credal Learning

Varun Venkatesh[1] **Eyke Hüllermeier**[2,3] **Bernd Bischl**[1,2] **Mina Rezaei**[1,2]

[1]Institute of Statistics, LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany,
[3]Institute of Informatics, LMU Munich, Germany

Abstract

Real-world learning tasks often encounter uncertainty due to covariate shift and noisy or inconsistent labels. However, existing robust learning methods merge these effects into a single distributional uncertainty set. In this work, we introduce a novel structured credal learning framework that explicitly separates these two sources. Specifically, we derive geometric bounds on the total variation diameter of structured credal sets and demonstrate how this quantity decomposes into contributions from covariate shift and expected label disagreement. This decomposition reveals a *gating effect*: covariate modulates how much label disagreement contributes to the joint uncertainty such that seemingly benign covariate shifts can substantially increase the effective uncertainty. We also establish finite-sample concentration bounds in a fixed covariate regime and demonstrate that this quantity can be efficiently estimated. Lastly, we show that robust optimization over these structured credal sets reduces to a tractable discrete min–max problem, avoiding ad-hoc robustness parameters. Overall, our approach provides a principled and practical foundation for robust learning under combined covariate and label mechanism ambiguity.

1 INTRODUCTION

Predictive uncertainty is crucial for developing reliable and trustworthy machine learning algorithms used to make critical decisions, such as medical diagnosis [Bordes et al., 2010, Vahidi et al., 2024, Wang et al., 2026b], drug discovery [Svensson et al., 2025], and autonomous driving [Chytas et al., 2024]. This uncertainty arises from multiple, fundamentally distinct sources: distributional mismatch and inherent target ambiguity [Hüllermeier and Beringer, 2005, Gal and Ghahramani, 2016]. Specifically, *covariate shift* occurs when the input distribution differs between training and test time while the conditional labeling mechanism distribution remains unchanged. We refer to *label mechanism ambiguity* for uncertainty in the conditional label distribution arising from measurement noise, annotator variability, or intrinsic class overlap, and it imposes a limit on achievable prediction accuracy [Bordes et al., 2010, Natarajan et al., 2013, Huang and Chong, 2023]. Crucially, these sources demand different operational remedies—distributional adaptation for the former, improved supervision or label reconciliation for the latter.

Most existing methods primarily address these challenges in isolation and typically do not decouple marginal and conditional uncertainty in robustness certificates. Domain adaptation [Chen et al., 2025, Tamang et al., 2025], importance weighting [Li et al., 2020, Sugiyama et al., 2007], invariant learning methods [Li et al., 2024, Nguyen et al., 2025a], and distributionally robust optimization (DRO) [Nguyen et al., 2025c, Jeong et al., 2025, Zhu et al., 2025, Nguyen et al., 2025b] have shown strong performance in mitigating covariate shift by aligning feature distributions across environments, typically under the assumption of stable labeling mechanisms. In contrast, research on label mechanism ambiguity focuses on robust training or uncertainty-aware prediction under fixed input distributions [Hüllermeier and Beringer, 2005, Hüllermeier, 2014, Cai et al., 2025b, Zhou et al., 2022]. Recent works [Li and Zhu, 2024, Caprio et al., 2024] attempt to diagnostically decompose prediction error into marginal and conditional components. For example, the credal learning framework of Caprio et al. [2024], grounded in imprecise probability theory [Augustin et al., 2014], replaces single distributions with convex sets of distributions (credal sets) to more faithfully represent uncertainty, limited knowledge, and data variability. However, it does not provide formal predictive robustness guarantees that can explicitly attribute observed errors to covariate shift versus label mechanism ambiguity.

In this paper, we propose a structured credal learning frame-

work that explicitly distinguishes the sources of uncertainty into covariate shift and label mechanism ambiguity. Instead of modeling distributional uncertainty as an unstructured set of joint distributions, we construct credal sets as convex combinations of product measures formed by pairing plausible covariate distributions with plausible labeling mechanisms making the resulting robustness certificate interpretable and operationally tied to semantically meaningful alternatives.

Taken together, our contributions are: (i) We introduce a structured credal learning framework that explicitly decomposes distributional uncertainty into covariate shift and label mechanism ambiguity. (ii) We derive explicit bounds on the total-variation diameter of structured credal sets and show that it decomposes along the same two sources of uncertainty linked by a gating interaction. This yields interpretable robust generalization bounds directly attributable to each source. (iii) In the fixed-covariate regime, we prove the diameter of the credal set equals the maximum pairwise disagreement rate — a directly measurable statistic estimable at parametric rate — and establish a minimax lower bound. (iv) We show that robust optimization over structured credal sets reduces to a finite min–max problem over extreme-points, with an intrinsic, observable robustness radius that requires no extrinsic hyperparameter tuning.

2 RELATED WORK

Our work bridges credal learning theory and DRO.

Credal learning Theory (CLT) [Caprio et al., 2024] extends statistical learning theory grounded on imprecise probability [Augustin et al., 2014]. Recent work leverages this perspective to derive finite-sample generalization guarantees whose tightness is controlled by the total variation diameter $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ of the credal set [Löhr et al., 2025, Javanmardi et al., 2024]. Existing CLT formulations [Shariatmadar et al., 2025, Wang et al., 2026a, Mubashar et al., 2025, Caprio et al., 2025], however, treat uncertainty at the level of the joint distribution over (X, Y), thereby conflating heterogeneity in the covariate marginal P_X with ambiguity in the conditional labeling mechanism $P_{Y|X}$. In contrast, we introduce a structured credal construction that factorizes uncertainty across covariate distributions and labeling mechanisms, enabling a geometric decomposition of the induced diameter into interpretable covariate-driven and label-driven components.

Distributionally robust optimization defines uncertainty sets as Wasserstein or f-divergence balls around the empirical distribution [Ben-Tal et al., 2013, Ai and Ren, 2024, Cai et al., 2025a, Zhu et al., 2021, Ehyaei et al., 2024], requiring specification of an extrinsic radius parameter and often producing worst-case distributions with limited semantic interpretability. In contrast, our uncertainty set is the convex hull of coherent generative "worlds" (covariate $\times$ labeling mechanism), leading to an intrinsic, data-driven robustness quantity via the credal diameter decomposition. This yields robustness certificates that are directly attributable to distinct sources of distributional variability and a finite extreme-point reduction of the robust objective.

3 STRUCTURED CREDAL LEARNING FRAMEWORK

3.1 PRELIMINARY: CREDAL LEARNING THEORY

Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label spaces, equipped with σ-algebras $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$. The joint space $\Omega := \mathcal{X} \times \mathcal{Y}$ carries the product σ-algebra $\mathcal{A}_\Omega := \mathcal{A}_{\mathcal{X}} \otimes \mathcal{A}_{\mathcal{Y}}$.

CLT generalizes empirical risk minimization (ERM) by replacing the assumption of a single data-generating distribution with a *credal set*, i.e., a convex set $\mathcal{P}$ of plausible probability distributions over $\mathcal{X} \times \mathcal{Y}$, establishing generalization guarantees across $\mathcal{P}$. This framework yields a powerful robustness certificate: for an ERM $\hat{h}$ trained on a source $P \in \mathcal{P}$, the generalization capability extends to any target $Q \in \mathcal{P}$ via the following bound:

$$\mathbb{P}\left[L_Q(\hat{h}) - L_P(h^*) \;\leq\; \varepsilon^*(\delta, n, \mathcal{H}) \;+\; \eta\right] \;\geq\; 1-\delta, \quad (1)$$

where $L_Q(\hat{h})$ is the expected population risk on the target distribution under the ERM hypothesis $\hat{h}$ and $L_P(h^*)$ is the optimal risk on the source. Here, ε^* encapsulates the learnable statistical error [1], while η represents the *robustness penalty* defined by the credal diameter, $\eta = \mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) = \sup_{P,Q\in\mathcal{P}} d_{\mathrm{TV}}(P, Q) = \sup_{A\in\mathcal{A}_\Omega} |P(A) - Q(A)|$.

However, CLT compresses all uncertainty sources—covariate shift and label mechanism ambiguity alike—into the single scalar $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$. This conflation obscures whether robustness penalties arise from distributional mismatch or intrinsic labeling uncertainty, precluding targeted interventions. Two settings may yield identical credal diameters yet require different remedies: high covariate shift with deterministic labels calls for domain adaptation, while stable features with ambiguous labels demand improved supervision.

This naturally leads to the notion of a *structured credal set*, which serves as the central object of analysis in the remainder of our work. We formalize a structured uncertainty model that captures variability in both the covariate distribution and the label mechanism ambiguity.

[1] Under the corresponding realizability and hypothesis class $\mathcal{H}$ complexity assumptions, Corollaries 4.2, 4.6, and 4.11 in Caprio et al. [2024] are unified by the formulation in Equation (1). In particular, ε^* recovers the respective error terms ϵ^*, ϵ^{**}, and ϵ^{***}

3.2 COVARIATE–LABELING UNCERTAINTY

We model uncertainty in the data-generating process as arising from two conceptually distinct sources: variation in the feature marginal (covariate shift), and variation in the labeling mechanism (labeling uncertainty). Operationally, this corresponds to a two-stage generative description: first sample $X \sim P_X$, then sample $Y \mid X \sim P_{Y|X}(\cdot \mid X)$.

Definition 3.1 (Feature and Label Distributions)**.** We define two finite collections of probability measures:

- Let $\mathcal{E} := \{P_X^{(i)}\}_{i=1}^{N_X}$ be a collection of probability distributions on $(\mathcal{X}, \mathcal{A}_\mathcal{X})$, representing plausible feature distributions or *environments* for short (e.g., different demographics or imaging equipments).
- Let $\mathcal{L} := \{P_{Y|X}^{(j)}\}_{j=1}^{N_Y}$ be a finite collection of conditional probability distributions from $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ to $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$, representing plausible labeling mechanisms or *labelers* for short (e.g. different annotators).

Each combination of an environment and a labeler defines a coherent "world" or data-generating process. We formalize this coupling via the joint product distribution.

Definition 3.2 (Joint Distribution)**.** For $(i, j) \in [N_X] \times [N_Y]$, define P_{ij} on $(\Omega, \mathcal{A}_\Omega)$ for $A \in \mathcal{A}_\mathcal{X}$, $B \in \mathcal{A}_\mathcal{Y}$ by

$$P_{ij}(A \times B) := \int_A P_{Y|X}^{(j)}(B \mid x)\, \mathrm{d}P_X^{(i)}(x),$$

When densities exist, $p_{ij}(x, y) = p_X^{(i)}(x) \cdot p_{Y|X}^{(j)}(y|x)$.

3.3 STRUCTURED CREDAL SET

Definition 3.3 (Structured Credal Set)**.** The *structured credal set* is defined as the convex hull of the joint distributions induced by all environment–labeler pairs:

$$\mathcal{P} \;:=\; \mathrm{Conv}\left(\{P_{ij}\}_{i\in[N_X],\, j\in[N_Y]}\right). \tag{2}$$

Remark 3.4 (Justification and Construction). The convex hull in Definition 3.3 admits a simple generative interpretation: if the environment–labeler pair (i, j) is chosen based on a fixed unknown mixing distribution π over $[N_X] \times [N_Y]$, the resulting data-generating distribution is $P_\pi = \sum_{i,j} \pi_{ij} P_{ij}$. Since π is unknown, the set of all such mixtures is exactly the $\mathrm{Conv}\left(\{P_{ij}\}\right)$. This construction imposes no adversarial assumption yet encodes uncertainty about which coherent generative world (or mixtures thereof) may be encountered at deployment. In settings like crowdsourcing and medical imaging, where each annotator/expert defines a labeler while different hospitals or patient populations define environments, the sets $\mathcal{E}$ and $\mathcal{L}$ arise naturally. When the generative process is parametric, these can be constructed by discretizing plausible parameter ranges.

Scope. This construction assumes a rectangular admissible set: any plausible environment can be meaningfully paired with any plausible labeling mechanism i.e. all pairings P_{ij} represent coherent deployment worlds rather than arbitrary perturbations. When environments and labelers are intrinsically coupled, e.g., site-specific annotation protocols, the appropriate object is a restricted set i.e. if admissible pairings are constrained by $\mathcal{A} \subset [N_X] \times [N_Y]$, one should instead use $\mathcal{P}_\mathcal{A} := \mathrm{Conv}\{P_{ij} : (i, j) \in \mathcal{A}\}$; the finite extreme-point reduction still applies, but the full covariate–label diameter attribution then applies to the rectangular envelope $\mathcal{P}$ and may be conservative for $\mathcal{P}_\mathcal{A}$.

Lemma 3.5 (Extremal Supremum Property)**.** *Let $\mathcal{P}$ be the structured credal set. Then:*

(i) For any continuous convex $f : \mathcal{P} \to \mathbb{R}$,

$$\sup_{Q\in\mathcal{P}} f(Q) = \max_{Q\in\mathrm{ex}(\mathcal{P})} f(Q).$$

(ii) For any continuous, separately convex $f : \mathcal{P}^2 \to \mathbb{R}$,

$$\sup_{(Q_1,Q_2)\in\mathcal{P}^2} f(Q_1, Q_2) = \max_{(Q_1,Q_2)\in\mathrm{ex}(\mathcal{P})^2} f(Q_1, Q_2).$$

4 THEORETICAL ANALYSIS: GEOMETRY OF THE STRUCTURED CREDAL SET

The diameter $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) := \sup_{P,Q\in\mathcal{P}} d_{\mathrm{TV}}(P, Q)$ governs robust generalization bounds by Equation (1). We show that this diameter decomposes into interpretable contributions from covariate shift and labeling uncertainty, with a particularly clean form when $N_X = 1$ (Section 5).

4.1 TOTAL VARIATION DECOMPOSITION

By Lemma 3.5 and the convexity of d_{TV}, the $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ is determined by the maximal distance between the extreme points in $\mathcal{P}$, i.e., any two distinct data-generating processes P_{ij} and $P_{i'j'}$ within our set.

In the following, we analyze three cases: (1) pure labeling uncertainty under a fixed covariate (fixed i), (2) pure covariate shift under a shared labeler (fixed j), and (3) the general case where both sources of uncertainty vary simultaneously. Detailed proofs for all propositions and theorems in this section are provided in the Appendix A.2

Proposition 4.1 (Label Disagreement Distance)**.** *Let $P_{ij}, P_{ij'}$ share the same environment marginal $P_X^{(i)}$. Then*

$$d_{\mathrm{TV}}(P_{ij}, P_{ij'}) = \mathbb{E}_{X\sim P_X^{(i)}}\left[d_{\mathrm{TV}}\big(P_{Y|X}^{(j)}(\cdot \mid X), P_{Y|X}^{(j')}(\cdot \mid X)\big)\right].$$

When the environment is fixed, the TV distance between two joint distributions reduces to the expected disagreement between the labelers.

Proposition 4.2 (Environment Shift Distance)**.** *Let P_{ij} and $P_{i'j}$ be joint distributions sharing the same labeler $P_{Y|X}^{(j)}$ but differing in their environment marginals $P_X^{(i)}$ and $P_X^{(i')}$. Then*

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j}) \;=\; d_{\mathrm{TV}}\big(P_X^{(i)}, P_X^{(i')}\big).$$

Interpretation. Under pure environment shift with a fixed labeler, the joint TV distance exactly equals the distance between the environment marginals. While the data processing inequality [Cover and Thomas, 2006] dictates a possible loss of distinguishability when observing only the label space $\mathcal{Y}$, we measure divergence in the fully observable joint space Ω. Because the invariant labeler merely redistributes probability mass across $\mathcal{Y}$ at each x, it does not dilute the inherent discrepancy of the underlying features.

We now combine the results from the fixed-environment and fixed-labeler regimes to bound the distance between two arbitrary worlds P_{ij} and $P_{i'j'}$ where both sources of uncertainty vary.

Theorem 4.3 (General Distance Bounds)**.** *Let $P_{ij}, P_{i'j'} \in \mathcal{P}$ with $i \neq i'$ and $j \neq j'$. Then:*

$$\max_{k\in\{i,i'\}} \left|\mathbb{E}_{P_X^{(k)}}\left[d_{\mathrm{TV}}\left(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}\right)\right] - d_{\mathrm{TV}}\left(P_X^{(i)}, P_X^{(i')}\right)\right|$$

$$\leq\; d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \;\leq$$

$$d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}) + \min_{k\in\{i,i'\}} \mathbb{E}_{P_X^{(k)}}\left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')})\right]$$

Interpretation. Theorem 4.3 reveals that the joint distributional divergence arises not merely as a sum of covariate shift and labeler disagreement, but through an interaction between them. The bounds decompose the distance between P_{ij} and $P_{i'j'}$ via specific interpolation paths creating a dependency structure we term the *Gating Effect*.

Intuitively, the environment determines which parts of the labeler disagreement landscape are exposed. The expectation $\mathbb{E}_{X\sim P_X^{(k)}}\left[d_{\mathrm{TV}}\left(P_{Y|X}^{(j)}(\cdot|X), P_{Y|X}^{(j')}(\cdot|X)\right)\right]$ is therefore an environment-weighted disagreement: large disagreement on a region $R \subseteq \mathcal{X}$ affects the joint space only when $P_X^{(k)}(R)$ is non-negligible. This behavior is visualized in Figure 1, where the joint TV distance spikes precisely when the covariate window places mass on the disagreement region, despite nearly constant covariate-only shift. Hence, deployment can reveal label disagreement that was effectively hidden under the training environment.

High-Dimensional Gating. Appendix Figure 2b extends this to $\mathbb{R}^3$ visualizing a trajectory through a disagreement nebula. Even small, constant covariate shifts can induce sharp spikes in joint distributional divergence, indicating that high-dimensional safety may be a narrow corridor rather than a continuous region.

Consequently, training under one environment may therefore underestimate the effective impact of labeler disagreement if deployment occurs in an environment that concentrates mass in regions of high disagreement. The structured credal framework makes this coupling explicit, enabling principled reasoning about generalization under simultaneous covariate shift and labeling uncertainty by explicitly taking combinations over all pairs when bounding the credal-diameter.

4.2 DIAMETER BOUNDS FOR PRODUCT CREDAL SETS

Definition 4.4 (Component Diameters)**.** For a structured credal set $\mathcal{P}$, we define:

$$\eta_X := \max_{i,i'\in[N_X]} d_{\mathrm{TV}}\big(P_X^{(i)}, P_X^{(i')}\big),$$

$$\bar{\eta}_{Y|X} := \sup_x \max_{j,j'\in[N_Y]} d_{\mathrm{TV}}\big(P_{Y|X}^{(j)}(\cdot|x), P_{Y|X}^{(j')}(\cdot|x)\big)$$

$$\eta_{Y|X}^{*} := \max_{i\in[N_X]} \max_{j,j'\in[N_Y]} \mathbb{E}_{P_X^{(i)}}\left[d_{\mathrm{TV}}\big(P_{Y|X}^{(j)}(\cdot|X), P_{Y|X}^{(j')}(\cdot|X)\big)\right]$$

The $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ is the key quantity governing the looseness of robust generalization bounds (Equation (1)). Using the decompositions in Section 4.1, we can bound the global diameter of $\mathcal{P}$ via Definition 4.4. These quantities capture distinct aspects of uncertainty: η_X measures the maximal divergence in covariate distributions, while $\{\eta_{Y|X}^{*}, \bar{\eta}_{Y|X}\}$ capture labeler disagreement at differing levels of aggregation. The interplay between these quantities determines the geometry of the credal set.

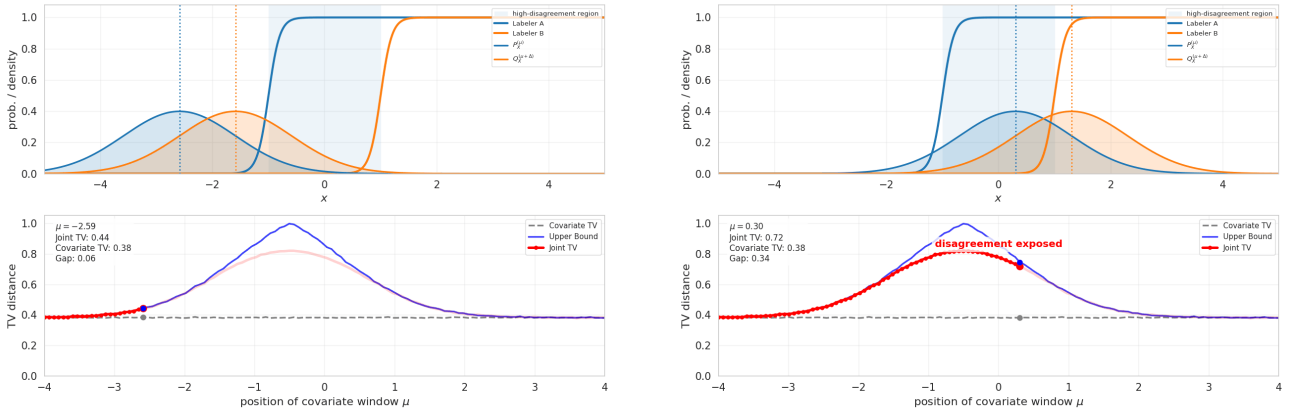
Theorem 4.5 (Diameter Decomposition)**.** *Let $\mathcal{P} = \mathrm{Conv}(\{P_{ij}\})$ be the structured credal set, then:*

$$\max\{\eta_X, \eta_{Y|X}^{*}\} \leq \mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \leq \eta_X + \eta_{Y|X}^{\mathrm{eff}},$$

where $\eta_{Y|X}^{\mathrm{eff}} := \min\{\eta_{Y|X}^{*},\, (1-\eta_X)\bar{\eta}_{Y|X}\}$.

Interpretation. The *lower bound* $\max\{\eta_X, \eta_{Y|X}^{*}\}$ establishes that distributional ambiguity cannot be reduced below the dominant source of uncertainty—whether from covariate shift or labeling uncertainty. The *upper bound* through the effective labeler disagreement, $\eta_{Y|X}^{\mathrm{eff}}$, shows that the diameter of the credal set is governed by an interaction between environments and labelers, operating through two mechanisms:

(i) Compounding regime ($\eta_{Y|X}^{\mathrm{eff}} = \eta_{Y|X}^{*}$): Covariate shift and labeling uncertainty compound additively. The binding



(a) Covariate shift outside the disagreement region.

(b) Covariate shift exposing the disagreement region.

Figure 1: **The Gating Effect.** We consider two sigmoid labeling mechanisms with thresholds at $x = \pm 1$, so that their disagreement is concentrated in the shaded region, and slide a pair of Gaussian covariate marginals across the input space while keeping their covariate-only distance fixed, $d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}) \approx 0.38$. In the left panel, most covariate mass lies outside the disagreement region, and the induced joint distance remains close to the covariate-shift baseline as the covariate shift occurs outside the disagreement region. In the right panel, the same magnitude of covariate shift places substantial mass on the disagreement region, making label-mechanism ambiguity visible in the joint distribution and increasing the joint TV distance. This demonstrates that the environment acts as a "gate", determining the visibility of labeler disagreement; seemingly benign covariate shifts can become catastrophic if they expose previously latent labeler disagreements. The red curve denotes the true joint TV distance along the trajectory, the gray dashed line the covariate-only TV, and the blue curve the upper bound from Theorem 4.3.

term is the environment-weighted disagreement: the relevant diameter-achieving pair places sufficient covariate mass on regions of labeler disagreement, so labeling uncertainty contributes through its full expected magnitude.

(ii) Overlap regime ($\eta_{Y|X}^{\mathrm{eff}} = (1-\eta_X)\bar{\eta}_{Y|X}$): Since feature distributions overlap in a region of total mass $1 - \eta_X$ (in TV for the worst-case pair), labeling uncertainty can contribute to the credal-diameter only on this overlapping mass. Outside the overlap, the uncertainty is already saturated by covariate shift.

Which term is active is determined by $\eta_{Y|X}^{\mathrm{eff}} = \min\{\eta_{Y|X}^{*}, (1-\eta_X)\bar{\eta}_{Y|X}\}$, i.e. whether the expected disagreement or the overlap-limited supremum disagreement is the binding constraint.

Remark 4.6 (Tightness). The lower bound is attained when the diameter-achieving pair $(P_{ij}, P_{i'j'})$ has either $i = i'$ (pure labeling uncertainty) or $j = j'$ (pure covariate shift). The upper bound is tight when covariate distributions have disjoint support ($\eta_X = 1$) or when label disagreement is constant across $\mathcal{X}$.

Connection to Generalization. Note that Theorem 4.5 yields a *factored generalization bound* in Equation (1):

$$\mathbb{P}\Big[L_Q(\hat{h}) - L_P(h^*) \;\le\; \varepsilon^*(\delta, n, \mathcal{H}) + \eta_X + \eta_{Y|X}^{\mathrm{eff}}\Big] \;\ge\; 1 - \delta\,, \tag{3}$$

decomposing the robustness penalty ($\eta = \mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$) into interpretable components ($\eta_X, \eta_{Y|X}^{\mathrm{eff}}$) allowing us to interpret the "safety budget" allocated to each source of uncertainty. While the general structured credal framework allows for simultaneous variations in environment and labelers, a critical sub-regime arises when the environment is stable, but the labeling mechanism is uncertain. This corresponds to the setting of $N_X = 1$, which we term the *Pure Labeling Uncertainty* regime.

5 THE PURE LABELING UNCERTAINTY REGIME

This regime deserves separate consideration for three reasons. First, it directly models practical scenarios such as **crowdsourcing**, where a fixed dataset is labeled by multiple annotators with varying expertise [Dawid and Skene, 1979, Raykar et al., 2010], and **inter-observer variability**, common in medical imaging where experts may grade the underlying pathology differently [McHugh, 2012, Mandrekar, 2011]. Second, the covariate shift term vanishes ($\eta_X = 0$), collapsing the bounded quantities in Theorem 4.5 – previously an abstract measure-theoretic quantity to a *directly measurable disagreement statistic* with an exact characterization. Third, this exactness enables finite-sample estimation and minimax optimality results providing a complete operational pipeline from data to robustness certificates.

Setting. Our analysis relies on the following structural assumptions formalized in Appendix B.2. We observe n i.i.d. inputs $X_1, \dots, X_n \sim P_X$ and, for each input X_i, labels from N_Y labeling mechanisms $\{P^{(k)}_{Y|X}\}_{k=1}^{N_Y}$. Given X_i, the labelers (deterministic or stochastic) are mutually independent and generate either (i) soft labels $\mathbf{p}_k(X_i) = P^{(k)}_{Y|X}(\cdot \mid X_i)$, or (ii) hard labels $Y_i^{(k)} \sim P^{(k)}_{Y|X}(\cdot \mid X_i)$ resulting in independent tuples $(X_i, Y_i^{(1)}, \dots, Y_i^{(N_Y)})$ across samples.

These assumptions capture the minimal structure of independent multi-annotator labeling on a fixed population, as in crowdsourcing or expert annotation, without assuming a latent ground truth or parametric noise model. They are required only for estimation and concentration, while the geometric characterization of the credal set and its diameter holds independently of these assumptions. Detailed proofs are provided in Appendix B

5.1 GEOMETRIC CHARACTERIZATION: DIAMETER AS DISAGREEMENT

Proposition 5.1 (Geometric Exactness of the Credal Diameter)**.** *Let $\mathcal{P} = \mathrm{Conv}(\{P_{1j}\}_{j=1}^{N_Y})$ be a structured credal set with fixed feature distribution P_X. The total-variation diameter is:*

$$\begin{aligned}\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) &= \eta^*_{Y|X} \\ &= \max_{j,j' \in [N_Y]} \mathbb{E}_{X \sim P_X}\left[d_{\mathrm{TV}}\big(P^{(j)}_{Y|X}, P^{(j')}_{Y|X}\big)\right].\end{aligned}$$

This diameter $\eta^*_{Y|X}$ bridges measure theory and practice, admitting concrete operational forms depending on the observation regime:

1. **Deterministic Hard Labels (Exact):** If the labelers are deterministic, the diameter is the *maximum pairwise disagreement rate*:

$$\eta^*_{Y|X} = \max_{j,j'} \mathbb{P}\left(f^{(j)}(X) \neq f^{(j')}(X)\right) \tag{4}$$

2. **Soft Label Regime (Exact):** If probability vectors $\mathbf{p}^{(j)}$ (e.g. confidence scores) are observed, the diameter is proportional to the expected L_1 distance:

$$\eta^*_{Y|X} = \tfrac{1}{2} \max_{j,j'} \mathbb{E}_X\left[\|\mathbf{p}^{(j)}(X) - \mathbf{p}^{(j')}(X)\|_1\right]. \tag{5}$$

3. **Stochastic Hard Labels (Conservative Bound):** If only realizations $y^{(j)}$ are observed,

$$\eta^*_{Y|X} \;\leq\; \max_{j,j'} \mathbb{P}\left(y^{(j)} \neq y^{(j')}\right). \tag{6}$$

   This ensures that using empirical disagreement to calibrate robustness is a *safe approximation* that never underestimates the necessary safety budget (See Remark 5.3).

4. **Noisy Labels (Aleatoric Uncertainty):** Under the binary symmetric noisy-annotator model with latent ground truth y^* and annotator error rates $\varepsilon_j \in [0, 0.5]$, the diameter of the resulting structured credal set admits a closed form:

$$\eta^*_{Y|X} = \max_{j,j'} \mathbb{P}(y^{(j)} \neq y^{(j')}) = \max_{j,j'} \big(\varepsilon_j + \varepsilon_{j'} - 2\varepsilon_j\varepsilon_{j'}\big). \tag{7}$$

   Consequently, if all annotators satisfy $\varepsilon_j \leq \varepsilon_{\max}$, the induced robustness radius is bounded by

$$\eta^*_{Y|X} \leq 2\varepsilon_{\max} - 2\varepsilon^2_{\max}, \tag{8}$$

   ensuring non-vacuous robustness even under weak supervision.

5.2 THE OBSERVABLE RADIUS: FROM HYPERPARAMETER TO STATISTICAL ESTIMATION

A key advantage of the structured credal framework in this regime is that the diameter of the uncertainty set is *observable* and *estimable* [2] directly from finite data.

Theorem 5.2 (Finite-Sample Concentration)**.** *Let $\hat{\eta}^*_{Y|X}$ be the empirical estimator of the diameter computed on a sample of size n. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\left|\hat{\eta}^*_{Y|X} - \eta^*_{Y|X}\right| \leq \sqrt{\frac{\ln\big(N_Y(N_Y - 1)/\delta\big)}{2n}}. \tag{9}$$

Interpretation and Complexity. The bound establishes that the diameter of the credal set can be estimated at the parametric rate $O(n^{-1/2})$ with only logarithmic dependence on the number of labelers N_Y. This logarithmic scaling is particularly favorable for crowdsourcing, where adding annotators improves coverage of the label space while incurring negligible sample complexity cost. For instance, with $N_Y = 10$ annotators, estimating the diameter to within $\epsilon = 0.1$ precision (95% confidence) requires only $n \approx 375$ samples.

Robustness to Assumptions. This estimation procedure: (i) does not specify error structures (e.g., confusion matrices); (ii) does not require ground truth labels; and (iii) assumes only conditional independence of annotators.

Remark 5.3 (Remark on Stochastic Hard Labels). In the stochastic hard label regime, the true TV diameter is non-identifiable as the underlying conditional distribution of the

[2] Note: We use *observable* to refer to sample-level quantities directly recorded (e.g., $X_{1:n}$ and $y^{(j)}_{1:n}$) or trivially computed (e.g., disagreements), and *estimable* for population quantities that can be estimated from these with finite-sample guarantees. In particular, $\eta^*_{Y|X}$ is a population diameter, while $\hat{\eta}^*_{Y|X}$ is its empirical estimator

labeler is non observable. However, since the TV distance is the infimum over all couplings, the observed *independent coupling* provides a mathematically valid upper bound. In this setting, the concentration bound in Eq. (9) applies to the *observed pairwise disagreement rate*. Since this upper-bounds the true diameter (Equation (6)), the finite-sample guarantee ensures the reliable estimation of a *conservative safety certificate*, effectively acting as a valid upper bound on the worst-case risk.

5.3 ROBUST GENERALIZATION AND OPTIMALITY

Finally, we connect the geometric diameter to learning guarantees. The following result establishes that $\eta^*_{Y|X}$ acts as a safety certificate, bounding the risk transfer between any two plausible interpretations of the data.

Corollary 5.4 (Robust Generalization)**.** *Let $\hat{h}$ be the empirical risk minimizer on a source distribution $P \in \mathcal{P}$. For any target $Q \in \mathcal{P}$ (representing an unknown true labeling mechanism), with probability $1-\delta$:*

$$L_Q(\hat{h}) - L_P(h^*) \;\le\; \underbrace{\varepsilon^*(\delta, n, \mathcal{H})}_{\text{statistical error}} + \underbrace{\eta^*_{Y|X}}_{\substack{\text{labeling uncertainty} \\ \text{(irreducible constant)}}} . \tag{10}$$

This decomposition cleanly separates two fundamentally different sources of error. The term ε^* captures the difficulty of learning the *source* distribution, vanishing as $n \to \infty$. In contrast, $\eta^*_{Y|X}$ captures the fundamental lack of consensus among labelers. It depends on the geometry of the credal set and cannot be reduced by collecting more data from the same ambiguous process. This distinction offers a *nuanced view of error*: a loss of magnitude $\eta^*_{Y|X}$ is not necessarily a model failure, but may be a reflection of genuine task ambiguity. Moreover, recent empirical studies with dense annotation consistently report substantial inter-annotator disagreement: in natural language inference, at least 40% disagreement is common for ambiguous examples [Jiang and de Marneffe, 2022], and majority-label instability reaches up to 31% when increasing annotation density [Nie et al., 2020]. In vision benchmarks, per-item agreement can drop to as low as 4%, corresponding to disagreement rates approaching 96% in extreme cases [Schmarje et al., 2022].

Is this ambiguity penalty tight? We show that no algorithm can circumvent this cost in the worst case.

Theorem 5.5 (Minimax lower bound)**.** *Fix* 0–1 *loss and assume $\mathcal{H}$ contains all measurable classifiers. For any $\eta \in (0,1)$, there exists a fixed covariate structured credal set $\mathcal{P} = \mathrm{Conv}(\{P_{1k}\}_{k=1}^{N_Y})$ with $\eta^*_{Y|X} = \mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ such that for every $n \in \mathbb{N}$,*

$$\inf_{\mathcal{A}} \sup_{P,Q \in \mathcal{P}} \mathbb{E}_{\mathcal{D}_n \sim P^{\otimes n}} \Big[L_Q(\mathcal{A}(\mathcal{D}_n)) - L_P^* \Big] \;\ge\; \frac{1}{2}\,\eta^*_{Y|X}. \tag{11}$$

Consequently, any upper bound with linear dependence on $\eta^*_{Y|X}$ is unimprovable in the worst case.

Theorem 5.5 confirms that $\eta^*_{Y|X}$ represents the *price of ambiguity*. To build intuition, consider two labelers with expected disagreement $\eta^*_{Y|X}$. On the region where they disagree, any hypothesis agreeing with one disagrees with the other. Any stochastic strategy incurs expected error at least 0.5 on the disagreement region.

Summary: A Three-Layer Unity. Together, Section 5 demonstrates that in the pure labeling uncertainty regime, the credal set is *fully observable*: its diameter is not a hyperparameter to be tuned but a statistic to be measured. $\eta^*_{Y|X}$ achieves a significant confluence in learning theory and plays a coherent role at three distinct levels of analysis. (i) **Geometrically**, it is the *diameter* of the uncertainty set (in TV distance). (ii) **Statistically**, it is an *observable parameter* (max pairwise disagreement) estimable at rate $O(n^{-1/2})$. (iii) **Theoretically**, it is the *minimax price of ambiguity*, establishing an irreducible floor on robust generalization. This unity transforms the $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ from an abstract robustness hyperparameter into a concrete, measurable, and fundamental limit of learning.

6 OPTIMIZATION

A central motivation for modeling distributional uncertainty is to optimize predictors against worst-case risk. Classical DRO formulates this objective as a minimax problem:

$$h^{\mathrm{DRO}} \;:=\; \arg\min_{h \in \mathcal{H}} \max_{Q \in \mathcal{U}_\rho(\hat{P})} L_Q(h), \tag{12}$$

where $\mathcal{U}_\rho(\hat{P})$ is typically a Wasserstein or f-divergence ball of radius ρ around the empirical distribution. While theoretically robust, this continuous formulation suffers from the *specification problem*—tuning the radius ρ is non-trivial, and worst-case perturbations often drift off the data manifold into non-semantic noise and the *computational problem* of optimizing over an infinite-dimensional space of probability measures, requiring nontrivial duality arguments or approximations [Duchi and Namkoong, 2021, Kuhn et al., 2019, 2025]. By defining the uncertainty set via the convex hull of coherent generative processes rather than an unstructured geometric ball, the intractable supremum collapses to a discrete, tractable, and interpretable min–max game.

6.1 REDUCTION TO EXTREME POINTS

Leveraging the geometric structure of $\mathcal{P}$ established in Definition 3.3, we derive an exact reduction of the optimization objective.

Proposition 6.1 (DRO Reduction)**.** *Let $\mathcal{P} = \mathrm{Conv}(\{P_{ij}\}_{i\in[N_X], j\in[N_Y]})$ be the structured credal*

set. The infinite-dimensional DRO problem in Equation (12) *reduces to a discrete minimax optimization over the extreme points:*

$$h^{\mathrm{DRO}} = \arg\min_{h\in\mathcal{H}} \max_{(i,j)\in[N_X]\times[N_Y]} L_{P_{ij}}(h). \qquad (13)$$

Significance. Proposition 6.1 reinterprets the optimization as Structured Group DRO [Sagawa et al., 2020], where groups are defined by generative factors (environment i, labeler j). More broadly, the finite game in (13) is related to robust risk minimization and minimax classification, where learning is also formulated as minimizing worst-case risk over a prescribed uncertainty set [Mazuelas et al., 2020]. In contrast, our uncertainty set is not generic but constructed from semantically interpretable environment–labeler worlds, so the resulting minimax game inherits the source-attributed geometry developed above. Crucially, this yields a **hyperparameter-free** framework: the robustness radius is intrinsic—determined by the $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \le \eta_X + \eta^{\mathrm{eff}}_{Y|X}$—rather than a tuned hyperparameter. The inner maximization admits a clear adversarial interpretation—Nature selects the worst-case combination of environment and labeler—while remaining constrained to semantically plausible generative worlds, avoiding the physically impossible perturbations possible in Wasserstein DRO.

In the pure labeling-uncertainty regime, the DRO objective in Proposition 6.1 reduces to enforcing robustness to the most challenging labeler under the current hypothesis, discouraging overfitting to majority opinion while guaranteeing performance on minority labelers—a property particularly important in medical imaging, where rare but critical findings may be identified by only a subset of expert radiologists and can be washed out by majority voting.

Algorithmic strategies. Because the inner maximization in Equation (13) is taken over a finite set of extreme points, the resulting minimax problem admits straightforward finite-world optimization via adversarial reweighting over structured worlds. In particular, the learner may be viewed as minimizing a weighted mixture of per-world risks, $\sum_{i,j} w_{ij}\,\hat{L}_{ij}(h)$, where w is the adversary's strategy over worlds (i,j). A natural hard-minimax strategy is greedy worst-world descent: at each iteration, identify the current worst world (i^*, j^*) and update using only its gradient (equivalently, a subgradient step on the pointwise maximum), which is computationally cheap but can yield non-smooth dynamics. To stabilize training, one can instead optimize the Log-Sum-Exp surrogate $\mathcal{L}_\tau(h) = \tau \log \sum_{(i,j)} \exp\!\left(\frac{\hat{L}_{ij}(h)}{\tau}\right)$ whose gradient is a weighted average of per-world gradients. As $\tau \downarrow 0$ this recovers hard minimax training; larger τ interpolates toward uniform averaging across worlds. However, this objective should be interpreted as an optional smoothing device rather than the exact hard-minimax problem. Alternatively, the exact finite problem may also be optimized directly by a primal–dual projected-gradient methods, avoiding finite-τ surrogate bias.

7 SYNTHETIC EXPERIMENTS

We also empirically validated the central geometric claims of the structured credal framework and assessed the statistical behavior of the proposed estimators on controlled synthetic families. These experiments are designed to verify the theoretical bounds in Sections 4 and 5 under known ground truth, where all quantities can be computed analytically.

Theorem 4.3 (General Distance Bounds.) First, we characterize the empirical tightness of the general distance decomposition by constructing a structured credal set using 20 Gaussian environments and 10 labelers, including soft (sigmoid) and deterministic hard (threshold) labelers with parameters sampled on uniform grids resulting in $40{,}000$ joint distribution pairs. Across all evaluated cases, the empirical TV distances satisfied the theoretical bounds with essentially zero slack (up to numerical precision) in the pure environment-shift and pure labeler-shift regimes and only moderate slack in the simultaneous joint-shift setting. Summarized results are shown in Table 1 with detailed breakdown in Table 2 in the Appendix.

Table 1: Tightness of general distance bounds (Theorem 4.3) in the joint-shift regime ($i \neq i'$, $j \neq j'$). Δ_{low} and Δ_{up} denote the mean gap between the true TV distance and the lower/upper bound, respectively.

Labeler Regime	$\Delta_{\mathrm{low}} \downarrow$	$\Delta_{\mathrm{up}} \downarrow$
Soft	0.242	0.165
Det. Hard	0.228	0.096

Disagreement-diameter characterization. We show that in the fixed-covariate regime the empirical diameter estimator is nearly unbiased for deterministic hard labelers (Table 3), remains accurate in the soft-label setting via the L_1-based form (Table 4), and continues to track the closed-form noisy-annotator prediction with low error across increasing noise budgets (Table 5).

Theorem 5.2 (Finite Sample Concentration & Labeler complexity). Finally, the absolute estimation error decays at the predicted $O(n^{-1/2})$ rate across multiple covariate distributions (Tables 6, 7, 8, 9 and 10; as well as Figure 3), with negligible empirical violation probability and stable tightness ratios, and we additionally study scaling with the number of labeling mechanisms, observing non-vacuous bounds and robustness even for large N_Y (Appendix D).

8 REAL-WORLD EXPERIMENTS

We perform additional experimental analysis on a real-world dataset based on human annotation disagreement. The goal is to test whether the fixed-covariate, pure-labeling-uncertainty regime yields an empirically useful and non-vacuous robustness certificate on real annotation data.

Dataset and experimental protocol We use the HS-Brexit dataset from the Learning with Disagreement (LeWiDi) challenge [Leonardelli et al., 2023], which contains $N = 1120$ Brexit-related tweets annotated by $N_Y = 6$ annotators. Each tweet is annotated with binary labels along four dimensions: hate speech, offensiveness, aggressiveness, and stereotyping. Our main experiments focus on the hate-speech label, while we also verify that the concentration behavior is qualitatively consistent across the remaining annotation dimensions. The official train/val/test splits contain $784/168/168$ examples. For the concentration experiment, we treat the merged dataset as an empirical population and repeatedly draw i.i.d. sub-samples from it. This allows us to test the observable-diameter estimator discussed in Proposition 5.1 and its finite-sample behavior from Theorem 5.2 under a controlled resampling protocol.

Observable disagreement diameter For deterministic hard labels, Proposition 5.1 shows that the pure-labeling-uncertainty diameter equals the maximum pairwise annotator disagreement rate,$\eta^\star_{Y|X} = \max_{j,j' \in [N_Y]} \mathbb{P}\Big(Y^{(j)} \neq Y^{(j')}\Big)$. On the hate-speech label, the empirical population diameter is $\eta^\star_{Y|X} = 0.218$, attained by annotators 2 and 5. Thus, the diameter is directly measurable from annotator disagreement, without tuning an external robustness radius. We estimate $\eta^\star_{Y|X}$ by the empirical maximum pairwise disagreement $\widehat{\eta}_{Y|X} = \max_{j,j'} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big\{Y_i^{(j)} \neq Y_i^{(j')}\Big\}$. Over 2000 i.i.d. resamples, the estimator concentrates at the predicted $O(n^{-1/2})$ rate. The median absolute error decreases from 0.082 at $n = 10$ to 0.009 at $n = 1000$. Moreover, the empirical 95th-percentile error remains below the Hoeffding union-bound envelope from Theorem 5.2 in all cases.

Risk-transfer certificates across annotators We next evaluate robust generalization on the empirical risk-transfer certificate suggested by Corollary 5.4. For each annotator P, we fine-tune a separate BERTweet classifier [Nguyen et al., 2020] on that annotator's training labels and evaluate the resulting model on the test labels of all six annotators. This produces a real-world test of transfer across plausible labeling mechanisms.

We instantiate the certificate using the empirical source generalization gap $\widehat{\epsilon}_P = \widehat{R}_{P,\mathrm{test}}(\widehat{h}) - \widehat{R}_{P,\mathrm{train}}(\widehat{h})$, and the observed population disagreement radius $\eta^\star_{Y|X} = 0.218$. Under the realizable-source proxy $L_P(h^\star) = 0$, the resulting empirical certificate takes the form $\widehat{R}_{Q,\mathrm{test}}(\widehat{h}) \leq \widehat{\epsilon}_P + \eta^\star_{Y|X}$. Aggregated over 10 random seeds, the certificate holds for all $36/36$ source-target annotator pairs. The worst observed target risk is 0.260, while the tightest slack is approximately 0.023. Hence, directly measured annotator disagreement provides a non-vacuous safety budget for transferring a classifier across plausible human labeling mechanisms.

Noisy-Label Finally, we evaluate a noisy-label stress test by treating one annotator as a reference labeler and simulating noisy annotators with flip probabilities $\{0.05,0.10,0.15,0.20,0.30\}$. Since the empirical quantity $\widehat{R}_{\mathrm{test}}(\widehat{h}) - \widehat{R}_{\mathrm{train}}(\widehat{h})$ can be negative in this setting; we use the conservative proxy $\widehat{\epsilon} = \max\Big\{0,\, \widehat{R}_{\mathrm{test}}(\widehat{h}) - \widehat{R}_{\mathrm{train}}(\widehat{h})\Big\}$. Across all source-target noisy-annotator pairs, the observed-disagreement certificate again upper-bounds the target risk. The certificate holds for all $25/25$ noisy source-target pairs, with minimum slack 0.06.

More details on real-world experiments have been provided in the Appendix E

9 CONCLUSION

We introduced a structured credal learning framework for decomposing the distributional uncertainty into covariate distributions and labeling mechanisms: By leveraging structure, we derived sharp geometric bounds on the diameter of the induced credal set and established an exact characterization in the pure label uncertainty regime. This result transforms robustness from a hyperparameter-dependent quantity into a directly measurable statistic governed by annotator disagreement. We further show that robust optimization over structured credal sets admits a tractable discrete formulation, replacing ad-hoc tuned robustness-radius with an intrinsic robustness quantity determined by the structured credal set yielding interpretable worst-case objectives. Together, these results provide a principled theoretical foundation for robust learning under simultaneous covariate shift and labeling uncertainty, bridging the gap between abstract robustness guarantees and empirically observable uncertainty sources.

Future Work Several directions follow from this work. First, extending the structured credal framework to continuous families of environments and annotators, potentially parameterized by latent variables, would broaden its applicability to large-scale deployment settings. Second, integrating the proposed robustness formulation with modern representation learning pipelines (e.g., self-supervised models) may enable joint optimization of feature invariance and label disagreement robustness. Another promising direction is the incorporation of task-specific structure, such as hierarchical labels or structured outputs, into the labeling mechanism component of the credal set. This may yield finer-grained robustness guarantees in complex prediction tasks.

Acknowledgements

We gratefully acknowledge the funding that supported this work. MR and BB were supported by the Amazon Research Award 2024. We also thank Vandita Shukla and Dr. Matthias Aßenmacher for their valuable feedback.

References

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 641–665. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ai24a.html.

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

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. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/23359484.

Antoine Bordes, Nicolas Usunier, and Jason Weston. Label ranking under ambiguous supervision for learning semantic correspondences. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 103–110, Haifa, Israel, June 2010. Omnipress. URL https://icml.cc/2010/papers/331.pdf.

Tiffany Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *Operations Research*, 74(2):898–916, 2025a. URL https://doi.org/10.1287/opre.2023.0217.

Zhongze Cai, Hansheng Jiang, and Xiaocheng Li. Out-of-distribution robust optimization. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 521–539. PMLR, 21–25 Jul 2025b. URL https://proceedings.mlr.press/v286/cai25c.html.

Michele Caprio, Maryam Sultana, Eleni G. Elia, and Fabio Cuzzolin. Credal learning theory. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 38665–38694. Curran Associates, Inc., 2024. doi: 10.52202/079017-1221.

Michele Caprio, Shireen K Manchingal, and Fabio Cuzzolin. Credal and interval deep evidential classifications. *arXiv preprint arXiv:2512.05526*, 2025.

Yuen Chen, Haozhe Si, Guojun Zhang, and Han Zhao. Moment alignment: Unifying gradient and hessian matching for domain generalization. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 705–736. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/chen25f.html.

Sotirios Panagiotis Chytas, Vishnu Lokhande, and Vikas Singh. Pooling image datasets with multiple covariate shift and imbalance. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 57282–57304, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346806.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021. doi: 10.1214/20-AOS2004. URL https://doi.org/10.1214/20-AOS2004.

Ahmad-Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. Wasserstein distributionally robust optimization through the lens of structural causal models and individual fairness. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 42430–42467. Curran Associates, Inc., 2024. doi: 10.52202/079017-1344.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

URL https://proceedings.mlr.press/v48/gal16.html.

Xia Huang and Kai Fong Ernest Chong. GenKL: An Iterative Framework for Resolving Label Ambiguity and Label Non-conformity in Web Images Via a New Generalized KL Divergence. *International Journal of Computer Vision*, 131(11):3035–3059, November 2023. ISSN 1573-1405. doi: 10.1007/s11263-023-01815-9. URL https://doi.org/10.1007/s11263-023-01815-9.

Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. In *International Symposium on Intelligent Data Analysis*, pages 168–179. Springer, 2005.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2013.09.003. URL https://www.sciencedirect.com/science/article/pii/S0888613X13001722. Special issue: Harnessing the information contained in low-quality data sources.

Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 116987–117014. Curran Associates, Inc., 2024. doi: 10.52202/079017-3714.

Jinyong Jeong, Hyungu Kahng, and Seoung Bum Kim. Multi-expert distributionally robust optimization for out-of-distribution generalization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 101081–101116. Curran Associates, Inc., 2025.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374, 2022. doi: 10.1162/tacl_a_00523. URL https://aclanthology.org/2022.tacl-1.78/.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019. URL https://doi.org/10.1287/educ.2019.0198.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025. doi: 10.1017/S0962492924000084.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.314. URL https://aclanthology.org/2023.semeval-1.314/.

Dong Li, Chen Zhao, Minglai Shao, and Wenjun Wang. Learning fair invariant representations under covariate and correlation shifts simultaneously. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1174–1183, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679727. URL https://doi.org/10.1145/3627673.3679727.

Fengpei Li, Henry Lam, and Siddharth Prusty. Robust importance weighting for covariate shift. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 352–362. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/li20b.html.

Mengting Li and Chuang Zhu. Noisy label processing for classification: A survey. *arXiv preprint arXiv:2404.04159*, 2024.

Timo Löhr, Paul Hofman, Felix Mohr, and Eyke Hüllermeier. Credal prediction based on relative likelihood. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 145184–145213. Curran Associates, Inc., 2025.

Jayawant N. Mandrekar. Measures of Interrater Agreement. *Journal of Thoracic Oncology*, 6(1):6–7, January 2011. ISSN 1556-0864. doi: 10.1097/JTO.0b013e318200f983. URL https://doi.org/10.1097/JTO.0b013e318200f983.

Santiago Mazuelas, Andrea Zanoni, and Aritz Pérez. Minimax classification with 0-1 loss and performance guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 302–312. Curran Associates, Inc., 2020.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Muhammad Mubashar, Shireen Kudukkil Manchingal, and Fabio Cuzzolin. Random-set large language models. *arXiv preprint arXiv:2504.18085*, 2025.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Anh T Nguyen, Lam Tran, Anh Tong, Tuan-Duy H Nguyen, and Toan Tran. Casual: Conditional support alignment for domain adaptation with label shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19668–19676, 2025a.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics, October 2020. doi: 10.18653/v1/2020.emnlp-demos.2. URL https://aclanthology.org/2020.emnlp-demos.2/.

Hai Dai Nguyen, Hiroshi Mamitsuka, and Atsuyoshi Nakamura. Multiple wasserstein gradient descent algorithm for multi-objective distributional optimization. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 3182–3199. PMLR, 21–25 Jul 2025b. URL https://proceedings.mlr.press/v286/nguyen25a.html.

Quan M. Nguyen, Nishant A Mehta, and Cristóbal A Guzmán. Beyond minimax rates in group distributionally robust optimization via a novel notion of sparsity. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46073–46115. PMLR, 13–19 Jul 2025c. URL https://proceedings.mlr.press/v267/nguyen25e.html.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734/.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. URL http://jmlr.org/papers/v11/raykar10a.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? - a data-centric image classification benchmark for noisy and ambiguous label estimation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33215–33232. Curran Associates, Inc., 2022.

Keivan Shariatmadar, Neil Yorke-Smith, Ahmad Osman, Fabio Cuzzolin, Hans Hallez, and David Moens. Generalized decision focused learning under imprecise uncertainty–theoretical study. *arXiv preprint arXiv:2502.17984*, 2025.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007. URL http://jmlr.org/papers/v8/sugiyama07a.html.

Emma Svensson, Hannah Rosa Friesacher, Susanne Winiwarter, Lewis Mervin, Adam Arany, and Ola Engkvist. Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artificial Intelligence in the Life Sciences*, 7:100128, 2025. ISSN 2667-3185. doi: https://doi.org/10.1016/j.ailsci.2025.100128. URL https://www.sciencedirect.com/science/article/pii/S2667318525000042.

Lakpa Tamang, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. Handling out-of-distribution data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):5948–5966, 2025. doi: 10.1109/TKDE.2025.3592614.

Amirhossein Vahidi, Simon Schosser, Lisa Wimmer, Yawei Li, Bernd Bischl, Eyke Hüllermeier, and Mina Rezaei. Probabilistic self-supervised representation learning via

scoring rules minimization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 34748–34766, 2024.

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, pages 26319–26327, 2026a.

Kun Wang, Jin Liu, Zeliang An, and Yuting Lu. Udel: Rethinking uncertainty dynamic estimation learning for ambiguous medical image segmentation. *Digital Signal Processing*, 169:105723, 2026b. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2025.105723. URL https://www.sciencedirect.com/science/article/pii/S1051200425007456.

Chen Zhou, Mohit Prabhushankar, and Ghassan AlRegib. On the ramifications of human label uncertainty. *arXiv preprint arXiv:2211.05871*, 2022.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 280–288. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/zhu21a.html.

Mingye Zhu, Yi Liu, Zheren Fu, Yongdong Zhang, and Zhendong Mao. Leveraging robust optimization for llm alignment under distribution shifts. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 171182–171208. Curran Associates, Inc., 2025.

Structured Credal Learning (Appendix)

Varun Venkatesh[1] **Eyke Hüllermeier**[2,3] **Bernd Bischl**[1,2] **Mina Rezaei**[1,2]

[1]Institute of Statistics, LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany,
[3]Institute of Informatics, LMU Munich, Germany

A PROOFS FROM SECTIONS 3 AND 4

A.1 LEMMA 3.5

Proof. Since $\mathcal{P}$ is the convex hull of finitely many probability measures $\{P_{ij}\}_{(i,j)\in[N_X]\times[N_Y]}$, it is a compact convex polytope and its finite set of extreme points satisfies $\mathrm{ex}(\mathcal{P}) \subseteq \{P_{ij}\}$.

Case (i) The function $f : \mathcal{P} \to \mathbb{R}$ is continuous and convex. By the Bauer Maximum Principle, $\sup_{Q\in\mathcal{P}} f(Q) = \max_{Q\in\mathrm{ex}(\mathcal{P})} f(Q)$.

Case (ii) Let $f\colon \mathcal{P}^2 \to \mathbb{R}$ be continuous and separately convex. Fix the second argument $Q_2 \in \mathcal{P}$. The partial map $Q_1 \mapsto f(Q_1, Q_2)$ is continuous and convex on $\mathcal{P}$. Applying the $k = 1$ case, it attains its supremum at an extreme point $Q_1^\star \in \mathrm{ex}(\mathcal{P})$. We define this partial supremum as:

$$g(Q_2) = \max_{Q_1\in\mathrm{ex}(\mathcal{P})} f(Q_1, Q_2).$$

Because $\mathrm{ex}(\mathcal{P})$ is a finite set, $g(Q_2)$ is the pointwise maximum of a finite number of functions, each of which is continuous and convex in Q_2. Since the pointwise maximum of finitely many continuous, convex functions is itself continuous and convex, $g\colon \mathcal{P} \to \mathbb{R}$ satisfies the conditions of the $k = 1$ case.

Applying the Bauer Maximum Principle to g over $Q_2 \in \mathcal{P}$, the supremum is attained at an extreme point $Q_2^\star \in \mathrm{ex}(\mathcal{P})$, yielding:

$$\sup_{(Q_1,Q_2)\in\mathcal{P}^2} f(Q_1, Q_2) = \sup_{Q_2\in\mathcal{P}} g(Q_2) = \max_{Q_2\in\mathrm{ex}(\mathcal{P})} g(Q_2) = \max_{(Q_1,Q_2)\in\mathrm{ex}(\mathcal{P})^2} f(Q_1, Q_2).$$

This completes the proof. □

A.2 PROOFS FOR PROPOSITIONS & THEOREMS 4.1 - 4.5

A.2.1 Proof for Proposition 4.1

Using the density representation with respect to product measure $\mu_\Omega = \mu_X \otimes \mu_Y$:

$$d_{\mathrm{TV}}(P_{ij}, P_{ij'}) = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{Y}} \left| p_X^{(i)}(x)p_{Y|X}^{(j)}(y|x) - p_X^{(i)}(x)p_{Y|X}^{(j')}(y|x) \right| d\mu_Y(y)\, d\mu_X(x).$$

Since $p_X^{(i)}(x) \geq 0 \,\forall\, x \in \mathcal{X}$ as it is a probability density, we can factor it out of the absolute value:

$$= \frac{1}{2}\int_{\mathcal{X}} p_X^{(i)}(x) \left(\int_{\mathcal{Y}} \left| p_{Y|X}^{(j)}(y|x) - p_{Y|X}^{(j')}(y|x) \right| d\mu_Y(y) \right) d\mu_X(x).$$

Recognizing that $\frac{1}{2}\int_{\mathcal{Y}} |p^{(j)}_{Y|X}(y|x) - p^{(j')}_{Y|X}(y|x)| d\mu_Y(y) = d_{\mathrm{TV}}(P^{(j)}_{Y|X}(\cdot|x), P^{(j')}_{Y|X}(\cdot|x))$:

$$= \int_{\mathcal{X}} p^{(i)}_X(x) \cdot d_{\mathrm{TV}}\left(P^{(j)}_{Y|X}(\cdot|x), P^{(j')}_{Y|X}(\cdot|x)\right) d\mu_X(x) = \mathbb{E}_{X\sim P^{(i)}_X}\left[d_{\mathrm{TV}}\left(P^{(j)}_{Y|X}(\cdot|X), P^{(j')}_{Y|X}(\cdot|X)\right)\right].$$

A.2.2 Proof for Proposition 4.2

Using the density representation with respect to product measure μ_Ω, the joint densities factorize as $p_{ij}(x,y) = p^{(i)}_X(x)p^{(j)}_{Y|X}(y|x)$ and $p_{i'j}(x,y) = p^{(i')}_X(x)p^{(j)}_{Y|X}(y|x)$.

Plugging this into the definition of TV distance:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j}) = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{Y}} \left|p^{(i)}_X(x)p^{(j)}_{Y|X}(y|x) - p^{(i')}_X(x)p^{(j)}_{Y|X}(y|x)\right| d\mu_Y(y)\, d\mu_X(x).$$

Since the conditional density $p^{(j)}_{Y|X}(y|x)$ is non-negative, we factor it out:

$$= \frac{1}{2}\int_{\mathcal{X}} \left|p^{(i)}_X(x) - p^{(i')}_X(x)\right| \left(\int_{\mathcal{Y}} p^{(j)}_{Y|X}(y|x)\, d\mu_Y(y)\right) d\mu_X(x).$$

As $\int_{\mathcal{Y}} p^{(j)}_{Y|X}(y|x)\, d\mu_Y(y) = 1$, substituting 1 in the above equation yields:

$$= \frac{1}{2}\int_{\mathcal{X}} \left|p^{(i)}_X(x) - p^{(i')}_X(x)\right| \cdot 1\, d\mu_X(x) \tag{14}$$

$$= d_{\mathrm{TV}}\left(P^{(i)}_X, P^{(i')}_X\right). \tag{15}$$

A.2.3 Proof for Theorem 4.3

Proof. We proceed through two decomposition paths and combine the resulting bounds.

Lower bound.

Step 1: Establishing exact intermediate distances.

By Proposition 4.1 applied with shared $P^{(i)}_X$ and $P^{(i')}_X$ respectively, let

$$d_{\mathrm{TV}}(P_{ij},\, P_{ij'}) = A_i, \qquad d_{\mathrm{TV}}(P_{i'j},\, P_{i'j'}) = A_{i'}. \tag{16}$$

By Proposition 4.2 (exact covariate-shift distance with shared labeler), applied with shared $P^{(j)}_{Y|X}$ and $P^{(j')}_{Y|X}$ respectively, let

$$d_{\mathrm{TV}}(P_{ij},\, P_{i'j}) = C, \qquad d_{\mathrm{TV}}(P_{ij'},\, P_{i'j'}) = C. \tag{17}$$

Step 2: Path A (change labeler first, then environment).

Consider the intermediate chain $P_{ij} \to P_{ij'} \to P_{i'j'}$. By the reverse triangle inequality for the total variation metric:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \;\geq\; \left|d_{\mathrm{TV}}(P_{ij}, P_{ij'}) - d_{\mathrm{TV}}(P_{ij'}, P_{i'j'})\right|. \tag{18}$$

Substituting the exact values from (16) and (17) into (18):

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \;\geq\; |A_i - C|. \tag{Path A}$$

Step 3: Path B (change environment first, then labeler).

Consider the intermediate chain $P_{ij} \to P_{i'j} \to P_{i'j'}$. By the reverse triangle inequality:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \;\geq\; |d_{\mathrm{TV}}(P_{ij}, P_{i'j}) - d_{\mathrm{TV}}(P_{i'j}, P_{i'j'})|\,. \tag{19}$$

Substituting the exact values from (16) and (17) into (19):

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \;\geq\; |C - A_{i'}| = |A_{i'} - C|\,. \tag{Path B}$$

Taking the maximum over both paths and substituting the definitions yields the stated bound.

Upper Bound. By the triangle inequality for total variation distance:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \leq d_{\mathrm{TV}}(P_{ij}, P_{i'j}) + d_{\mathrm{TV}}(P_{i'j}, P_{i'j'}).$$

Applying Proposition 4.2 to the first term:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j}) = d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}).$$

Applying Proposition 4.1 to the second term:

$$d_{\mathrm{TV}}(P_{i'j}, P_{i'j'}) = \mathbb{E}_{X \sim P_X^{(i')}} \left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}) \right].$$

Alternatively, using the path through $P_{ij'}$:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \leq d_{\mathrm{TV}}(P_{ij}, P_{ij'}) + d_{\mathrm{TV}}(P_{ij'}, P_{i'j'}),$$

which gives:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \leq \mathbb{E}_{X \sim P_X^{(i)}} \left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}) \right] + d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}).$$

Taking the minimum over both paths and substituting the definitions yields the stated bound. □

A.2.4 Proof for Theorem 4.5

Definition A.1 (Pointwise Label Disagreement)**.** For each $x \in \mathcal{X}$, the *pointwise label disagreement* is:

$$\eta_{Y|X}(x) := \max_{j,j' \in [N_Y]} d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}). \tag{20}$$

Definition A.2 (Maximum Expected Label Disagreement under $P_X^{(i)}$)**.** The *max expected label disagreement under* $P_X^{(i)}$ is:

$$\eta_{Y|X}^{(i)} := \max_{j,j' \in [N_Y]} \mathbb{E}_{X \sim P_X^{(i)}} \left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}) \right]. \tag{21}$$

Lower Bound 1: $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \eta_X$**.**

The diameter is achieved at extreme points. Consider any pair (i, i') with $i \neq i'$ and fix any $j \in [N_Y]$:

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq d_{\mathrm{TV}}(P_{ij}, P_{i'j}) = d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}),$$

where the equality follows from Proposition 4.2. Taking the maximum over all pairs (i, i'):

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \max_{i,i' \in [N_X]} d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')}) = \eta_X.$$

Lower Bound 2: $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \eta_{Y|X}^{*}$

Fix any $i \in [N_X]$ and consider pairs (j, j') with $j \neq j'$:

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq d_{\mathrm{TV}}(P_{ij}, P_{ij'}) = \mathbb{E}_{X \sim P_X^{(i)}} \left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}) \right],$$

where the equality follows from Proposition 4.1. Taking the maximum over (j, j'):

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \max_{j,j'} \mathbb{E}_{X \sim P_X^{(i)}} \left[d_{\mathrm{TV}}(P_{Y|X}^{(j)}, P_{Y|X}^{(j')}) \right] = \eta_{Y|X}^{(i)}.$$

Since this holds for all i, taking the maximum over i:

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \max_{i \in [N_X]} \eta_{Y|X}^{(i)} = \eta_{Y|X}^*$$

Combining both lower bounds:

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \geq \max\{\eta_X, \eta_{Y|X}^*\}.$$

Upper Bound 1: $\mathrm{diam}_{TV}(\mathcal{P}) \leq \eta_X + \eta_{Y|X}^*$**.**

Let P_{ij} and $P_{i'j'}$ be arbitrary extreme points of $\mathcal{P}$. By the triangle inequality,

$$d_{TV}(P_{ij}, P_{i'j'}) \ \leq \ d_{TV}(P_{ij}, P_{i'j}) + d_{TV}(P_{i'j}, P_{i'j'}).$$

For the first term, since the conditional distribution is shared, Proposition 4.2 yields

$$d_{TV}(P_{ij}, P_{i'j}) \ \leq \ d_{TV}(P_X^{(i)}, P_X^{(i')}) \ \leq \ \eta_X.$$

For the second term, the marginal distribution is shared, and by Proposition 4.1,

$$d_{TV}(P_{i'j}, P_{i'j'}) = \mathbb{E}_{X \sim P_X^{(i')}} \left[d_{TV} \left(P_{Y|X}^{(j)}(\cdot|X), P_{Y|X}^{(j')}(\cdot|X) \right) \right] \ \leq \ \eta_{Y|X}^*.$$

Combining the two bounds gives us:

$$d_{TV}(P_{ij}, P_{i'j'}) \ \leq \ \eta_X + \eta_{Y|X}^*.$$

Since the supremum of total variation over a convex set is attained at extreme points, the result follows.

Upper Bound 2: $\mathrm{diam}_{TV}(\mathcal{P}) \leq \eta_X + (1 - \eta_X)\bar{\eta}_{Y|X}$**.**

Using the density representation with respect to product measure μ_Ω, We define the "overlap" density $m(x)$ and the "excess" densities $\Delta_i(x)$ and $\Delta_{i'}(x)$ as follows:

$$m(x) := \min\{p_X^{(i)}(x), p_X^{(i')}(x)\}, \qquad \Delta_i(x) := p_X^{(i)}(x) - m(x), \qquad \Delta_{i'}(x) := p_X^{(i')}(x) - m(x)$$

By construction $\Delta_i(x) \cdot \Delta_{i'}(x) = 0$ for all x, and $\int_{\mathcal{X}} m(x)\, d\mu_X(x) = 1 - d_{\mathrm{TV}}(P_X^{(i)}, P_X^{(i')})$. We decompose the Total Variation distance using these densities:

$$\begin{aligned} d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) &= \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} |p_i(x) p_{Y|X}^{(j)}(y|x) - p_{i'}(x) p_{Y|X}^{(j')}(y|x)|\, d\mu_Y(y)\, d\mu_X(x) \\ &= \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} |(m(x) + \Delta_i(x)) p_{Y|X}^{(j)} - (m(x) + \Delta_{i'}(x)) p_{Y|X}^{(j')}|\, d\mu_Y(y)\, d\mu_X(x) \end{aligned}$$

Applying the triangle inequality:

$$\begin{aligned} d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \leq & \underbrace{\frac{1}{2} \int_{\mathcal{X}} m(x) \int_{\mathcal{Y}} |p_{Y|X}^{(j)} - p_{Y|X}^{(j')}|\, d\mu_Y(y)\, d\mu_X(x)}_{\text{Term I}} \\ & + \underbrace{\frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} |\Delta_i(x) p_{Y|X}^{(j)} - \Delta_{i'}(x) p_{Y|X}^{(j')}|\, d\mu_Y(y)\, d\mu_X(x)}_{\text{Term II}} \end{aligned}$$

Analyzing Term I (Label Disagreement on Overlap): We regroup the factor of $1/2$ into the inner integral to identify the conditional Total Variation distance:

$$\begin{aligned}
\text{Term I} &= \int_{\mathcal{X}} m(x) \underbrace{\left(\frac{1}{2} \int_{\mathcal{Y}} |p^{(j)}_{Y|X}(y|x) - p^{(j')}_{Y|X}(y|x)| \, d\mu_Y(y) \right)}_{d_{\mathrm{TV}}(P^{(j)}_{Y|X}(\cdot|x), P^{(j')}_{Y|X}(\cdot|x))} d\mu_X(x) \\
&= \int_{\mathcal{X}} m(x) d_{\mathrm{TV}}(P^{(j)}_{Y|X}(\cdot|x), P^{(j')}_{Y|X}(\cdot|x)) \, d\mu_X(x) \\
&\leq \int_{\mathcal{X}} m(x) \max_{j,j'} d_{\mathrm{TV}}(P^{(j)}_{Y|X}(\cdot|x), P^{(j')}_{Y|X}(\cdot|x)) \, d\mu_X(x) \\
&= \int_{\mathcal{X}} m(x) \eta_{Y|X}(x) \, d\mu_X(x) \\
&\leq \int_{\mathcal{X}} m(x) \sup_{x \in \mathcal{X}} \eta_{Y|X}(x) \, d\mu_X(x) \\
&= \int_{\mathcal{X}} m(x) \bar{\eta}_{Y|X} \, d\mu_X(x) \\
&= \bar{\eta}_{Y|X}(1 - d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X))
\end{aligned}$$

Analyzing Term II (Feature Mismatch):

1. **Step 1: Understanding the Disjoint Supports:** For any specific input x, only one of three cases can occur:
   - **Case A:** $p^{(i)}_X(x) > p^{(i')}_X(x)$. Here, the minimum is $p^{(i')}_X(x)$. Therefore, $\Delta_{i'}(x) = 0$ and $\Delta_i(x) > 0$.
   - **Case B:** $p^{(i')}_X(x) > p^{(i)}_X(x)$. Here, the minimum is $p^{(i)}_X(x)$. Therefore, $\Delta_i(x) = 0$ and $\Delta_{i'}(x) > 0$.
   - **Case C:** $p^{(i')}_X(x) = p^{(i)}_X(x)$. Here, $\Delta_i(x) = \Delta_{i'}(x) = 0$.

   Crucially, this means the product $\Delta_i(x) \cdot \Delta_{i'}(x)$ is always zero. They never exist simultaneously at the same point x.
2. **Step 2: Why the Absolute Value Vanishes** Consider the expression inside the absolute value brackets:

   $$\left| \Delta_i(x) \cdot p^{(j)}_{Y|X}(y|x) - \Delta_{i'}(x) \cdot p^{(j')}_{Y|X}(y|x) \right|$$

   Using the property established above:
   - In **Case A** (where $\Delta_{i'} = 0$), the term becomes $|\Delta_i \cdot p^{(j)}_{Y|X}(y|x) - 0| = \Delta_i \, p^{(j)}_{Y|X}(y|x)$.
   - In **Case B** (where $\Delta_i = 0$), the term becomes $|0 - \Delta_{i'} p^{(j')}| = \Delta_{i'} \cdot \, p^{(j')}_{Y|X}(y|x)$.
   - In **Case C**, (where $\Delta_i = \Delta_{i'} = 0$), the term becomes $|0 - 0| = 0$

   In all three cases, the result is mathematically identical to the *sum* of the two terms (since at least one term is always zero, adding it changes nothing). Thus, we can drop the absolute value and simply add them:

   $$\left| \Delta_i(x) \cdot p^{(j)}_{Y|X}(x) - \Delta_{i'}(x) \cdot p^{(j')}_{Y|X}(x) \right| = \Delta_i(x) \cdot p^{(j)}_{Y|X}(y|x) + \Delta_{i'}(x) \cdot p^{(j')}_{Y|X}(y|x)$$

3. **Step 3: Splitting the Integral:** We substitute this sum back into the main integral:

   $$\text{Term II} = \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(\Delta_i(x) p^{(j)}_{Y|X}(y|x) + \Delta_{i'}(x) p^{(j')}_{Y|X}(y|x) \right) d\mu_Y(x) \, d\mu_X(x)$$

   By the linearity of integration, we can separate this into two distinct integrals:

   $$= \frac{1}{2} \left[\int_{\mathcal{X}} \Delta_i(x) \left(\int_{\mathcal{Y}} p^{(j)}_{Y|X}(y|x) d\mu_Y(y) \right) d\mu_X(x) + \int_{\mathcal{X}} \Delta_{i'}(x) \left(\int_{\mathcal{Y}} p^{(j')}_{Y|X}(y|x) d\mu_Y(y) \right) d\mu_X(x) \right]$$

4. **Integrating out y:** Since $p_{Y|X}$ is a probability distribution:

$$\int_{\mathcal{Y}} p^{(j)}_{Y|X}(y|x)\, d\mu_Y(x) \leq 1 \quad \text{and} \quad \int_{\mathcal{Y}} p^{(j')}_{Y|X}(y|x)\, d\mu_Y(x) \leq 1$$

Substituting these 1s into the equation simplifies it to integrals over just x:

$$\leq \frac{1}{2}\left[\int_{\mathcal{X}} \Delta_i(x) \cdot 1\, d\mu_X(x) + \int_{\mathcal{X}} \Delta_{i'}(x) \cdot 1\, d\mu_X(x)\right]$$

5. **Final Evaluation:** The integral of the excess density $\int \Delta_i(x) d\mu_X(x)$ represents the total probability mass unique to distribution P_i. A standard identity of Total Variation distance states that $\int \Delta_i(x) d\mu_X(x) = d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X)$.

$$= \frac{1}{2}\left[d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X) + d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X)\right]$$

$$= d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X)$$

Combining terms, for any pair of distributions:

$$d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) \leq d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X) + (1 - d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X))\bar{\eta}_{Y|X}.$$

Let $\delta = d_{\mathrm{TV}}(P^{(i)}_X, P^{(i')}_X)$. The function $f(\delta) = \delta + (1-\delta)\bar{\eta}_{Y|X}$ is monotonically increasing in δ (since $\bar{\eta}_{Y|X} \leq 1$). Thus, we maximize the bound by taking the maximum feature distance η_X:

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) \leq \eta_X + (1 - \eta_X)\bar{\eta}_{Y|X}.$$

Combined Upper Bound: $\mathrm{diam}_{TV}(\mathcal{P}) \leq \eta_X + \min\{\eta^*_{Y|X}, (1-\eta_X)\bar{\eta}_{Y|X}\}$

Because $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ admits both quantities as upper bounds, taking their minimum and applying the identity $\min(a + b, a + c) = a + \min(b, c)$ yields

$$\eta_X + \min\Big\{\eta^*_{Y|X},\ (1-\eta_X)\bar{\eta}_{Y|X}\Big\},$$

which establishes the claimed bound.

B PROOFS FOR $N_X = 1$ REGIME

We consider the *pure label uncertainty* regime, where $N_X = 1$. In this case, the covariate distribution is fixed and unique, denoted by $P^{(1)}_X = P_X$, and therefore the covariate diameter satisfies $\eta_X = 0$.

B.1 PROPOSITION 5.1: GEOMETRIC EXACTNESS

By Proposition 4.1, for any two joint distributions P_{1j} and $P_{1j'}$ in the structured credal set that share the same marginal feature distribution P_X but differ in their conditional label distributions, the total variation distance admits the exact representation

$$d_{\mathrm{TV}}(P_{1j}, P_{1j'}) = \mathbb{E}_{X\sim P_X}\left[d_{\mathrm{TV}}\Big(P^{(j)}_{Y|X}(\cdot \mid X), P^{(j')}_{Y|X}(\cdot \mid X)\Big)\right].$$

The diameter of the credal set $\mathcal{P}$ is defined as the maximum total variation distance between any two of its elements. Since $\mathcal{P}$ is the convex hull of the extreme points $\{P_{1j}\}_{j=1}^{N_Y}$ and total variation distance is maximized at extreme points, we obtain

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) = \max_{j,j'} d_{\mathrm{TV}}(P_{1j}, P_{1j'}).$$

Substituting the expression above yields

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) = \max_{j,j'} \mathbb{E}_{X\sim P_X}\left[d_{\mathrm{TV}}\Big(P^{(j)}_{Y|X}(\cdot \mid X), P^{(j')}_{Y|X}(\cdot \mid X)\Big)\right] = \eta^*_{Y|X},$$

which completes the proof.

B.2 THEOREM 5.2: FINITE SAMPLE CONCENTRATION

To derive concentration bounds for the diameter of the structured credal set as defined in Definition 3.3 in the fixed-feature regime ($N_X = 1$), we must first formalize the data-generating process involving multiple labeling mechanisms. Let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the label space. We consider a single feature distribution P_X and a set of N_Y distinct labeling mechanisms $\{P_{Y|X}^{(k)}\}_{k=1}^{N_Y}$.

We posit the following assumptions regarding the sampling and annotation process:

Assumption B.1 (i.i.d. Feature Sampling)**.** The feature vectors are drawn independently and identically distributed from the marginal feature distribution:

$$X_1, \ldots, X_n \sim_{i.i.d.} P_X$$

Assumption B.2 (Labeling Mechanism)**.** Each annotator $k \in \{1, \ldots, K\}$ is characterized by a conditional distribution $P_{Y|X}^{(k)}$ over the label space $\mathcal{Y} = \{1, \ldots, C\}$. For each sample $i \in \{1, \ldots, n\}$ and input realization $X_i = x_i$:

(a) The annotator's *belief* or *confidence scores* is given by the probability vector:

$$\mathbf{p}_k(x_i) := \left(P_{Y|X}^{(k)}(1 \mid x_i), \ldots, P_{Y|X}^{(k)}(C \mid x_i)\right) \in \Delta^{C-1}$$

where Δ^{C-1} is the $(C-1)$-dimensional probability simplex.

(b) The *realized label* $y_i^{(k)} \in \mathcal{Y}$ is drawn according to:

$$y_i^{(k)} \mid X_i = x_i \sim P_{Y|X}^{(k)}(\cdot \mid x_i)$$

Assumption B.3 (Observation Model)**.** We observe one of the following, depending on the regime:

(a) **Soft Label Regime:** The full probability vector $\mathbf{p}_k(x_i) \in \Delta^{C-1}$ is observed for each sample i and labeling mechanism k.

(b) **Hard Label Regime:** Only the realized label $y_i^{(k)} \in \mathcal{Y}$ is observed for each sample i and labeling mechanism k.

Assumption B.4 (Structure of Labeling Mechanism)**.** The conditional distributions $\{P_{Y|X}^{(k)}\}_{k=1}^{K}$ satisfy one of the following structural conditions:

(a) **Deterministic:** For each annotator k, there exists a measurable function $f_k : \mathcal{X} \to \mathcal{Y}$ such that:

$$P_{Y|X}^{(k)}(c \mid x) = \mathbb{I}[f_k(x) = c] \quad \forall c \in \mathcal{Y},\ x \in \mathcal{X}$$

Equivalently, the conditional probability vector

$$\mathbf{p}_k(x) := \left(P_{Y|X}^{(k)}(1 \mid x), \ldots, P_{Y|X}^{(k)}(C \mid x)\right)$$

belongs to $\{e_1, \ldots, e_C\}$, where $e_c \in \mathbb{R}^C$ denotes the c-th standard basis vector (one-hot encoding).

(b) **Stochastic:** The conditional distributions are non-degenerate, i.e., there exists at least one $k \in [N_Y]$ and $x \in \mathcal{X}$ such that:

$$P_{Y|X}^{(k)}(c \mid x) \in (0, 1) \quad \text{for some } c \in \mathcal{Y}$$

Assumption B.5 (Conditional Independence of Labelers)**.** Conditional on the feature realization X_i, the labelers operate independently. That is, for any $j \neq k$:

$$y_i^{(j)} \perp\!\!\!\perp y_i^{(k)} \mid X_i$$

This implies that the joint conditional distribution of all N_Y labels factorizes:

$$P(y_i^{(1)}, \ldots, y_i^{(K)} \mid X_i) = \prod_{k=1}^{N_Y} P_{Y|X}^{(k)}(y_i^{(k)} \mid X_i)$$

Assumption B.6 (Sample Independence)**.** The generation processes for distinct samples $i \neq j$ are mutually independent. Specifically, the tuple $\mathbf{Z}_i = (X_i, y_i^{(1)}, \ldots, y_i^{(N_Y)})$ is independent of $\mathbf{Z}_j$.

Establishing that the fully observed tuples are i.i.d. allows us to apply standard concentration inequalities (e.g., Hoeffding's Inequality) to the estimation of pairwise disagreement rates.

Lemma B.7 (i.i.d. Structure of Annotated Tuples)**.** *Under Assumptions B.1–B.6, the joint tuples* $\mathbf{Z}_i = (X_i, y_i^{(1)}, \ldots, y_i^{(K)})$ *for* $i = 1, \ldots, n$ *are independent and identically distributed random variables taking values in the product space* $\Omega_{joint} = \mathcal{X} \times \mathcal{Y}^K$.

Proof. Let μ be the base measure on $\mathcal{X}$ and ν be the base measure on $\mathcal{Y}$. We analyze the joint probability density (or mass) function of a single tuple $\mathbf{Z}_i$, denoted as $p_{\mathbf{Z}}(x, y^{(1)}, \ldots, y^{(K)})$.

By the definition of conditional probability and Assumption B.2, the joint density can be decomposed as:

$$p_{\mathbf{Z}}(x, y^{(1)}, \ldots, y^{(K)}) = p_X(x) \cdot p(y^{(1)}, \ldots, y^{(K)} \mid x)$$

Applying Assumption B.5 (Conditional Independence), the conditional term factorizes:

$$p(y^{(1)}, \ldots, y^{(K)} \mid x) = \prod_{k=1}^{K} p_{Y|X}^{(k)}(y^{(k)} \mid x)$$

Thus, the joint distribution for any index i is uniquely determined by the fixed distributions P_X and $\{P_{Y|X}^{(k)}\}_{k=1}^K$:

$$p_{\mathbf{Z}}(x, y^{(1)}, \ldots, y^{(K)}) = p_X(x) \prod_{k=1}^{K} p_{Y|X}^{(k)}(y^{(k)} \mid x) \tag{22}$$

Since P_X and $\{P_{Y|X}^{(k)}\}$ do not depend on the index i, every tuple $\mathbf{Z}_i$ follows the exact same distribution defined in Eq. (22). Furthermore, by Assumption B.6, the tuples are mutually independent.

Therefore, the sequence $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ consists of independent and identically distributed draws from the joint distribution $P_{\mathbf{Z}}$. □

Definition B.8 (Labeling Regimes)**.** The combination of Assumptions B.3 and B.4 defines three distinct regimes:

Regime	**Observation**	**Structure**
Soft Labels	$\mathbf{p}_k(x_i) \in \Delta^{C-1}$	-
Deterministic Hard Labels	$y_i^{(k)} \in \mathcal{Y}$	Deterministic
Stochastic Hard Labels	$y_i^{(k)} \in \mathcal{Y}$	Stochastic

In the soft-label regime, we observe the conditional distributions themselves. For a deterministic hard-label regime, the observed label can be mapped to a conditional probability vector (refer to Assumption B.4). For a classification task with C classes, for each input x_i, the labeling mechanism k provides a probability vector $\mathbf{p}_k(x_i) \in \Delta^{C-1}$, where $\mathbf{p}_k(x_i) \equiv P_{Y|X}^{(k)}(\cdot \mid x_i)$.

Definition B.9 (Empirical TV Distance)**.** The Total Variation distance between two probability vectors $\mathbf{p}, \mathbf{q}$ is defined as $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) := \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$. For annotators j and k, the empirical TV distance based on n samples is:

$$\hat{d}_{TV}^{jk} = \frac{1}{n} \sum_{i=1}^{n} d_{TV}(\mathbf{p}_j(x_i), \mathbf{p}_k(x_i)). \tag{23}$$

Definition B.10 (Population TV Distance)**.** The true population disagreement is defined as the expected TV distance over the feature distribution P_X:

$$d_{TV}^{\star,jk} = \mathbb{E}_{x \sim P_X}\left[d_{TV}(\mathbf{p}_j(x), \mathbf{p}_k(x))\right]. \tag{24}$$

We define the maximum disagreement over all pairs (j, k) as:

$$\hat{\eta}_{Y|X} = \max_{j,k} \hat{d}_{TV}^{jk}, \quad \eta_{Y|X}^{\star} = \max_{j,k} d_{TV}^{\star,jk}$$

Theorem B.11 (Concentration Bound)**.** *Let N_Y be the number of labeling mechanisms providing either deterministic hard labels or soft labels. For any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$|\hat{\eta}_{Y|X} - \eta^{\star}_{Y|X}| \leq \sqrt{\frac{\ln(N_Y(N_Y-1)/\delta)}{2n}}. \tag{25}$$

Proof. We establish this result through analysis of the concentration of the empirical mean of the pairwise TV distances.

Step 1: Define the random variables.

For a fixed pair of annotators (j,k), we define the random variable corresponding to the disagreement on the i-th sample:

$$Z_i := d_{TV}(\mathbf{p}_j(x_i), \mathbf{p}_k(x_i)) = \frac{1}{2}\|\mathbf{p}_j(x_i) - \mathbf{p}_k(x_i)\|_1.$$

We analyze the properties of Z_i:

1. **Boundedness:** By the definition of Total Variation distance on probability space, $Z_i \in [0,1]$ for any valid probability vectors.
2. **i.i.d.:** The annotator policies $\mathbf{p}_j(\cdot)$ and $\mathbf{p}_k(\cdot)$ are fixed, deterministic functions given x_i. Since the inputs $\{x_i\}_{i=1}^n$ are drawn i.i.d. from P_X (Under Assumptions B.1–B.6), the transformations $Z_i = g(x_i)$ are also i.i.d.
3. **Expectation:** By linearity of expectation and the definition of the population parameter:

$$\mathbb{E}[Z_i] = \mathbb{E}_{x\sim P_X}[d_{TV}(\mathbf{p}_j(x), \mathbf{p}_k(x))] = d_{TV}^{\star,jk}.$$

Step 2: Apply Hoeffding's inequality. For bounded i.i.d. random variables $Z_i \in [0,1]$ with mean $\mu = d_{TV}^{\star,jk}$, Hoeffding's inequality states:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2). \tag{26}$$

Substituting our quantities:

$$\mathbb{P}\left(\left|\hat{d}_{TV}^{jk} - d_{TV}^{\star,jk}\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2). \tag{27}$$

Step 3: Apply union bound over all pairs.

There are $M = \binom{N_Y}{2} = \frac{N_Y(N_Y-1)}{2}$ distinct pairs of annotators. We apply the union bound to control the probability that *any* pair exceeds the error threshold ϵ.

Let E_{jk} be the event $\{|\hat{d}_{TV}^{jk} - d_{TV}^{\star,jk}| > \epsilon\}$.

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{(j,k)} E_{jk}\right) &\leq \sum_{(j,k)} \mathbb{P}(E_{jk}) \\
&\leq M \cdot 2\exp(-2n\epsilon^2) \\
&= \frac{N_Y(N_Y-1)}{2} \cdot 2\exp(-2n\epsilon^2) \\
&= N_Y(N_Y-1)\exp(-2n\epsilon^2).
\end{aligned} \tag{28}$$

We set this total probability to δ and solve for ϵ:

$$N_Y(N_Y-1)\exp(-2n\epsilon^2) = \delta \tag{29}$$

$$\exp(-2n\epsilon^2) = \frac{\delta}{N_Y(N_Y-1)} \tag{30}$$

$$-2n\epsilon^2 = \ln\left(\frac{\delta}{N_Y(N_Y-1)}\right) N_Y \tag{31}$$

$$\epsilon = \sqrt{\frac{\ln(N_Y(N_Y-1)/\delta)}{2n}}. \tag{32}$$

Step 4: Bound on the maxima.

With probability at least $1-\delta$, the bound ϵ holds for all pairs (j,k) simultaneously. Using the property that for any two sequences, $|\max_i a_i - \max_i b_i| \leq \max_i |a_i - b_i|$, we obtain:

$$\begin{aligned} |\hat{\eta}_{Y|X} - \eta^{\star}_{Y|X}| &= \left| \max_{j,k} \hat{d}_{TV}^{jk} - \max_{j,k} d_{TV}^{\star,jk} \right| \\ &\leq \max_{j,k} \left| \hat{d}_{TV}^{jk} - d_{TV}^{\star,jk} \right| \\ &\leq \epsilon \\ &= \sqrt{\frac{\ln(N_Y(N_Y-1)/\delta)}{2n}}. \end{aligned} \tag{33}$$

□

B.3 COROLLARY 5.4: ROBUST GENERALIZATION

Credal Learning Theory provides a high-probability bound on the expected risk of the empirical risk minimizer $\hat{h}$ when evaluated under a possibly misspecified distribution $Q \in \mathcal{P}$, distinct from the distribution $P \in \mathcal{P}$ from which the training data were drawn. In particular, CLT controls the excess risk incurred due to distributional ambiguity within the credal set.

Specifically, with probability at least $1-\delta$, the following robustness bound holds:

$$L_Q(\hat{h}) - L_P(h^*) \;\leq\; \epsilon^*(\delta, n, \mathcal{H}) \;+\; \mathrm{diam}_{\mathrm{TV}}(\mathcal{P}),$$

where $\epsilon^*(\delta, n, \mathcal{H})$ denotes the statistical estimation error, depending on the confidence level δ, the sample size n, and the hypothesis class $\mathcal{H}$, and $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P})$ is the total variation diameter of the credal set.

Within the structured credal framework, in the pure label uncertainty regime ($N_X = 1$), the diameter admits the exact characterization

$$\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) = \eta^*_{Y|X}.$$

Substituting this expression into the bound above yields the desired result.

B.3.1 Theorem 5.5: Minimax Optimality

Proof. Fix an arbitrary sample size $n \in \mathbb{N}$ and an arbitrary learning algorithm $\mathcal{A}$ mapping datasets $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ to (possibly randomized) predictors.

Step 1: Construct a $N_X = 1$ credal set with diameter $\eta^*_{Y|X} = \eta$.

Let $(\mathcal{X}, \mathcal{A}_X, P_X)$ be any probability space and fix any $\eta \in (0,1)$. Choose a measurable set $S \in \mathcal{A}_X$ such that $P_X(S) = \eta$. Define two *deterministic* labeling mechanisms (hard labels)

$$f_1(x) \equiv 0, \qquad f_2(x) := \mathbf{1}\{x \in S\}.$$

Equivalently, define two conditional label distributions

$$P^{(1)}_{Y|X}(\cdot \mid x) = \delta_0, \qquad P^{(2)}_{Y|X}(\cdot \mid x) = \delta_{f_2(x)}.$$

Let P_{11} and P_{12} be the corresponding joint distributions (Definition 3.2):

$$P_{1j}(A \times B) = \int_A P^{(j)}_{Y|X}(B \mid x)\, dP_X(x), \qquad j \in \{1,2\}.$$

Finally, define the (structured) credal set

$$\mathcal{P} := \mathrm{Conv}\big(\{P_{11}, P_{12}\}\big).$$

We compute its fixed-feature diameter $\eta^*_{Y|X}$. For $x \notin S$, $P^{(1)}_{Y|X}(\cdot \mid x) = P^{(2)}_{Y|X}(\cdot \mid x) = \delta_0$ so the TV distance is 0; for $x \in S$, we have δ_0 vs. δ_1 so the TV distance is 1. Hence

$$\eta^*_{Y|X} = \mathbb{E}_{X\sim P_X}\Big[d_{\mathrm{TV}}\big(P^{(1)}_{Y|X}(\cdot \mid X), P^{(2)}_{Y|X}(\cdot \mid X)\big)\Big] = \mathbb{E}_{X\sim P_X}\big[\mathbf{1}\{X \in S\}\big] = P_X(S) = \eta.$$

Thus $\mathrm{diam}_{\mathrm{TV}}(\mathcal{P}) = \eta^*_{Y|X} = \eta$ in this construction (cf. Proposition 5.1).

Step 2: Represent an arbitrary output of the learner and compute risks.

Given a dataset $\mathcal{D}_n$, let $\widehat{h}_{\mathcal{D}_n} := \mathcal{A}(\mathcal{D}_n)$. Allow $\widehat{h}_{\mathcal{D}_n}$ to be randomized; define the induced prediction probability

$$\pi_{\mathcal{D}_n}(x) \;:=\; \mathbb{P}\big(\widehat{h}_{\mathcal{D}_n}(x) = 1\big) \in [0,1], \qquad \mathbb{P}\big(\widehat{h}_{\mathcal{D}_n}(x) = 0\big) = 1 - \pi_{\mathcal{D}_n}(x).$$

Under P_{11} the true label is $Y = f_1(X) = 0$ almost surely, so the (conditional) pointwise misclassification probability at x equals $\pi_{\mathcal{D}_n}(x)$. Therefore

$$L_{P_{11}}(\widehat{h}_{\mathcal{D}_n}) = \mathbb{E}_{X\sim P_X}\big[\pi_{\mathcal{D}_n}(X)\big]. \tag{34}$$

Under P_{12} the true label is $Y = f_2(X) = \mathbf{1}\{X \in S\}$ almost surely, hence

$$\mathbb{P}\big(\widehat{h}_{\mathcal{D}_n}(X) \neq Y \mid X = x\big) = \begin{cases} \pi_{\mathcal{D}_n}(x), & x \notin S \;\; (Y = 0), \\ 1 - \pi_{\mathcal{D}_n}(x), & x \in S \;\; (Y = 1). \end{cases}$$

Therefore

$$L_{P_{12}}(\widehat{h}_{\mathcal{D}_n}) = \mathbb{E}_{X\sim P_X}\Big[\pi_{\mathcal{D}_n}(X)\mathbf{1}\{X \notin S\} + (1 - \pi_{\mathcal{D}_n}(X))\mathbf{1}\{X \in S\}\Big]. \tag{35}$$

Step 3: A two-point testing lower bound

Add (34) and (35). For each fixed dataset $\mathcal{D}_n$,

$$\begin{aligned} L_{P_{11}}(\widehat{h}_{\mathcal{D}_n}) + L_{P_{12}}(\widehat{h}_{\mathcal{D}_n}) &= \mathbb{E}_{X\sim P_X}\Big[\pi_{\mathcal{D}_n}(X) + \pi_{\mathcal{D}_n}(X)\mathbf{1}\{X \notin S\} + (1 - \pi_{\mathcal{D}_n}(X))\mathbf{1}\{X \in S\}\Big] \\ &= \mathbb{E}_{X\sim P_X}\Big[2\pi_{\mathcal{D}_n}(X)\mathbf{1}\{X \notin S\} + \big(\pi_{\mathcal{D}_n}(X) + 1 - \pi_{\mathcal{D}_n}(X)\big)\mathbf{1}\{X \in S\}\Big] \\ &= \mathbb{E}_{X\sim P_X}\Big[2\pi_{\mathcal{D}_n}(X)\mathbf{1}\{X \notin S\} + \mathbf{1}\{X \in S\}\Big] \\ &\geq \mathbb{E}_{X\sim P_X}\big[\mathbf{1}\{X \in S\}\big] = P_X(S) = \eta. \end{aligned}$$

Thus, for *every* dataset $\mathcal{D}_n$,

$$L_{P_{11}}(\widehat{h}_{\mathcal{D}_n}) + L_{P_{12}}(\widehat{h}_{\mathcal{D}_n}) \;\geq\; \eta. \tag{36}$$

Step 4: Take expectations over training data and lower bound the worst-case test risk.

Now choose the training distribution $P := P_{11} \in \mathcal{P}$. Taking expectations of (36) over $\mathcal{D}_n \sim P_{11}^{\otimes n}$ yields

$$\mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}\big[L_{P_{11}}(\widehat{h}_{\mathcal{D}_n})\big] + \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}\big[L_{P_{12}}(\widehat{h}_{\mathcal{D}_n})\big] \;\geq\; \eta.$$

Hence, the larger of the two expectations is at least half:

$$\max\Big\{\mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}[L_{P_{11}}(\widehat{h}_{\mathcal{D}_n})],\; \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}[L_{P_{12}}(\widehat{h}_{\mathcal{D}_n})]\Big\} \;\geq\; \frac{\eta}{2}. \tag{37}$$

Since $\{P_{11}, P_{12}\} \subseteq \mathcal{P}$, we have

$$\sup_{Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}\big[L_Q(\widehat{h}_{\mathcal{D}_n})\big] \;\geq\; \max\Big\{\mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}[L_{P_{11}}(\widehat{h}_{\mathcal{D}_n})],\; \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}[L_{P_{12}}(\widehat{h}_{\mathcal{D}_n})]\Big\} \;\geq\; \frac{\eta}{2},$$

where the last inequality is (37).

Step 5: Insert the baseline L^*_P and conclude the minimax lower bound. Because P_{11} is realizable by the deterministic classifier $h_1(x) \equiv 0$ (which we may assume lies in $\mathcal{H}$), we have $L^*_{P_{11}} = 0$. Therefore

$$\sup_{Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}\Big[L_Q(\widehat{h}_{\mathcal{D}_n}) - L^*_{P_{11}}\Big] = \sup_{Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P_{11}^{\otimes n}}\big[L_Q(\widehat{h}_{\mathcal{D}_n})\big] \;\geq\; \frac{\eta}{2}.$$

Since $P_{11} \in \mathcal{P}$, the supremum over $P \in \mathcal{P}$ can only increase the value:

$$\sup_{P,Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\Big[L_Q(\mathcal{A}(\mathcal{D}_n)) - L_P^*\Big] \;\geq\; \frac{\eta}{2}.$$

Finally, the above bound holds for *every* learning algorithm $\mathcal{A}$, hence taking $\inf_{\mathcal{A}}$ preserves it:

$$\inf_{\mathcal{A}} \sup_{P,Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\Big[L_Q(\mathcal{A}(\mathcal{D}_n)) - L_P^*\Big] \;\geq\; \frac{\eta}{2}.$$

Recalling that $\eta = \eta^*_{Y|X}$ in this construction, we obtain exactly

$$\inf_{\mathcal{A}} \sup_{P,Q\in\mathcal{P}} \mathbb{E}_{\mathcal{D}_n\sim P^{\otimes n}}\Big[L_Q(\mathcal{A}(\mathcal{D}_n)) - L_P^*\Big] \;\geq\; \frac{1}{2}\,\eta^*_{Y|X},$$

which proves the theorem. □

B.4 NOISY ANNOTATOR MODEL (ALEATORIC UNCERTAINTY)

We specialize the general structured credal learning framework to the classical noisy-annotator setting, where label uncertainty arises from irreducible annotation noise rather than model misspecification. This instantiation yields closed-form expressions for the induced conditional ambiguity radius, enabling direct interpretation of robustness guarantees in terms of annotator error rates.

- **Model (Binary Symmetric Noisy Annotators):** Let $Y = \{0, 1\}$ and assume a latent ground-truth label $y^* \in \{0, 1\}$. Annotator $j \in \{1, \ldots, K\}$ outputs $y^{(j)}$ through a binary symmetric channel with error rate $\varepsilon_j \in [0, 0.5]$:

  $$\mathbb{P}\Big(y^{(j)} = y^* \mid y^*\Big) = 1 - \varepsilon_j, \qquad \mathbb{P}\Big(y^{(j)} \neq y^* \mid y^*\Big) = \varepsilon_j.$$

  Assume conditional independence of annotators given y^*.

- **Annotator–Annotator Disagreement (Closed Form):** For any pair (j, j'), disagreement occurs iff exactly one annotator flips y^*. Thus,

  $$\mathbb{P}\Big(y^{(j)} \neq y^{(j')}\Big) = \mathbb{P}\Big(y^{(j)} = y^*,\, y^{(j')} \neq y^*\Big) + \mathbb{P}\Big(y^{(j)} \neq y^*,\, y^{(j')} = y^*\Big) = (1-\varepsilon_j)\varepsilon_{j'} + \varepsilon_j(1-\varepsilon_{j'}) = \varepsilon_j + \varepsilon_{j'} - 2\varepsilon_j\varepsilon_{j'}.$$

- **Credal Diameter / Robustness Radius:** In the fixed-covariate regime ($N_X = 1$), the structured credal diameter equals the maximum pairwise disagreement across annotators:

  $$\eta^*_{Y|X} = \max_{j,j'} \mathbb{P}\Big(y^{(j)} \neq y^{(j')}\Big) = \max_{j,j'} \big(\varepsilon_j + \varepsilon_{j'} - 2\varepsilon_j\varepsilon_{j'}\big).$$

- **Non-vacuous Upper Bound under Weak Supervision:** If all annotators satisfy $\varepsilon_j \leq \varepsilon_{\max}$, then

  $$\eta^*_{Y|X} \leq \max_{j,j'} \big(\varepsilon_{\max} + \varepsilon_{\max} - 2\varepsilon^2_{\max}\big) = 2\varepsilon_{\max} - 2\varepsilon^2_{\max}.$$

  In particular, for $\varepsilon_{\max} \leq 1/2$, this bound is strictly below 1, yielding a non-vacuous ambiguity radius even with noisy supervision.

Proof. The disagreement identity follows conditional independence given y^* and a two-term partition on which the annotator flips. The upper bound is immediate by monotonicity in each ε_j over $[0, 0.5]$. □

C PROOF FOR PROPOSITION 6.1

Proof. For any fixed hypothesis $h \in \mathcal{H}$, the expected risk $L_P(h) = \mathbb{E}_P[\ell((x, y), h)]$ is a strictly linear functional with respect to the probability measure P. With respect to the bounded loss-function ($\ell \in [0, 1]$), the mapping $P \mapsto L_P(h)$ is continuous.

Since $L_P(h)$ is continuous and linear (and thus convex) on the compact structured credal set $\mathcal{P}$, we can invoke Lemma 3.5 with $k = 1$. Thus the supremum risk is attained at an extreme point:

$$\sup_{P \in \mathcal{P}} L_P(h) = \max_{P \in \mathrm{ex}(\mathcal{P})} L_P(h).$$

By construction, $\mathcal{P}$ is the convex hull of the finite collection $\{P_{ij}\}$, meaning its extreme points are necessarily a subset of these generators ($\mathrm{ex}(\mathcal{P}) \subseteq \{P_{ij}\}$). Consequently, evaluating the maximum over the extreme points is equivalent to maximizing over the discrete set of indices $(i, j) \in [N_X] \times [N_Y]$.

Substituting this finite inner maximum back into the original outer minimization over $\mathcal{H}$ directly yields the discrete minimax formulation in (13). □

D ADDITIONAL EMPIRICAL ANALYSIS, ABLATION STUDIES, DISCUSSION, LIMITATIONS, AND FUTURE WORKS

D.1 EMPIRICAL ANALYSIS AND ABLATION STUDIES

Theorem 4.3 (General Distance Bounds.) We simulated a structured credal set by defining 20 Gaussian covariate distributions P_X, considering 15 distributions with standard deviation $\sigma = 1$ and means placed on a uniform grid over $[-3, 3]$, together with 5 additional randomized Gaussians. We further defined 10 labeling mechanisms. Soft labeling mechanisms were specified via conditional distributions $P_{Y|X}$ parameterized by sigmoid link functions with slope 1.0, while deterministic hard labeling mechanisms were implemented using threshold classifiers of the form $\mathbf{1}(x > \theta)$. Thresholds (and sigmoid biases) were selected from a grid over $[-4, 4]$. This construction yielded a total of $20 \times 20 \times 10 \times 10 = 40{,}000$ distribution pairs for evaluation. Table 2 presents a detailed breakdown of the tightness of the bounds across various settings. Across all evaluated cases, the empirical total variation distances satisfied the theoretical bounds with essentially zero slack (up to numerical precision) in the pure environment-shift and pure labeler-shift regimes and only moderate slack in the simultaneous joint-shift setting.

Disagreement-Diameter Characterization. We validated the observable diameter characterizations across both the deterministic hard labeling regime (Equation (4)), and the soft label regime (5). We sampled 500 random Gaussian distributions with varying means and variances and applied two distinct labeling mechanisms: (1) a threshold classifier (deterministic), (2) Probit link functions with $\kappa \in \{1, 3\}$. Across both settings, the empirical total (mean disagreement rate or mean L_1 distance) proved to be unbiased and tightly concentrated around the theoretical credal diameter. In the soft label regime, the mean estimation gap was negligible (3.65×10^{-5}), and deviations from the ground truth diameter remained within 0.05 for 99.9% of the trials. Together, these results demonstrate that the diameter of our structured credal set is fully observable in the fixed covariate regime and collapses the abstract geometric and measure-theoretic concept to simply measuring disagreements. The detailed metrics and quantiles are shown in Table 2. Finally, we also evaluated the noisy labeling mechanism framework (Equation (7) where the labeling mechanisms flipped the ground truth. According to Table 5, when allowing noise rates for the labeling mechanisms up to $\epsilon = 0.5$, the estimator remained robust, recovering the theoretical diameter with negligible bias ≈ 0.02, confirming that the diameter of the credal set can be accurately observed even when the labeling mechanisms are imperfect.

Theorem 5.2 (Sample Complexity). The empirical results also provide a strong validation for the concentration bounds derived in Theorem 5.2 across both deterministic hard labels and probabilistic soft labels. Over three covariate distributions ($\mathcal{N}(0, 1), \mathcal{N}(2, 1), \mathcal{N}(0, 2)$ in the hard-label regime, the empirical diameter converges at the theoretically consistent rate $O(\frac{1}{\sqrt{n}})$ with negligible violation probability and the tightness ratio between the theoretical bound and the 95^{th} percentile error remains stable around ≈ 2 indicating a practically informative nound at moderate sample sizes. In the soft label regime,

Table 2: Ablation study characterizing the tightness of the general distance bounds in Theorem 4.3. We evaluate 20 Gaussian covariates and 10 labeling mechanisms across all pair classes. The lower-gap and upper-gap are defined as $\epsilon_{\mathrm{LB}} := d_{\mathrm{TV}}(P_{ij}, P_{i'j'}) - \mathrm{LB}(P_{ij}, P_{i'j'})$ and $\epsilon_{\mathrm{UB}} := \mathrm{UB}(P_{ij}, P_{i'j'}) - d_{\mathrm{TV}}(P_{ij}, P_{i'j'})$, respectively. Entries report the mean gap with $[\min, \max]$ over all pairs within each class. All evaluated pairs satisfy the bounds, resulting in a 100% empirical coverage. Slack values on the order of 10^{-8} or smaller (numerical precision) and are reported as 0.

Pair class	#Pairs	ϵ_{LB} (mean [min,max])	ϵ_{UB} (mean [min,max])
Soft labels (Sigmoid)			
Joint Shift	34,200	0.242 [3.95e-04, 0.96]	0.165 [2.81e-04, 0.803]
Fixed Covariate (Conditional Shift only)	1,800	0 [0, 0]	0 [0, 0]
Fixed Labeling Mechanism (Covariate Shift only)	3,800	0 [0, 0]	0 [0, 0]
Identical pair	200	0 [0, 0]	0 [0, 0]
Hard labels (Thresholds)			
Joint Shift	34,200	0.228 [0, 1]	0.096 [0, 0.936]
Fixed Covariate (Conditional Shift only)	1,800	0 [0, 0]	0 [0, 0]
Fixed Labeling Mechanism (Covariate Shift only)	3,800	0 [0, 0]	0 [0, 0]
Identical pair	200	0 [0, 0]	0 [0, 0]

the convergence rate for a probabilistic labeling mechanism using both the probit link function and the sigmoid function for the covariate distribution $\mathcal{N}(0, 1)$ was theoretically consistent, and violations were absent. Moreover, in the soft label regime, the mean empirical diameter already tracks close to the population value at small n, reflecting the lower variance signal induced by observing the probabilities themselves.

Theorem 5.2 (Labeling Mechanism Complexity). We evaluate the stability of the diameter estimation as the number of labeling mechanisms N_Y scales from 2 to 1000. Since the true diameter of the structured credal set in the fixed covariate regime $N_X = 1$ is given by $\eta^*_{Y|X} = \mathbb{E}_{X \sim P_X}\left[d_{\mathrm{TV}}\left(P^{(j)}_{Y|X}, P^{(j')}_{Y|X}\right)\right]$, the *gating effect* plays an important role in determining the diameter of the credal set. To rigorously isolate the impact of mechanism complexity, we constructed a framework where the feature space $\mathcal{X}$ is partitioned into disjoint blocks defined by the quantiles of the marginal distribution P_X. Each labeling mechanism is restricted to a specific block, ensuring that any pair of mechanisms from different blocks has disjoint supports. Therefore, any pair of labeling mechanisms across blocks could be the maximal disagreeing pair. By fixing the probability mass of these blocks, we structurally pin the $\eta^*_{Y|X}$ at a constant value, preventing it from growing with N_Y. Since the blocks are defined via quantiles, this construction renders error bounds and theoretical properties invariant to the specific P_X, needing us to only evaluate the effects on a single feature distribution. Additionally, to also consider the possibility that adding a new labeling mechanism might also include an "outlier," we evaluated the mechanism complexity where $\eta^*_{Y|X}$ expands with the number of labeling mechanisms as well. Under both scenarios, the bounds remain consistently non-vacuous, with estimation error scaling logarithmically in N_Y. Crucially, the tightness ratio stabilizes between 1.4 and 3.3 even up to $N_Y = 1000$, confirming that the framework scales robustly to a high number of labeling mechanisms without degradation.

D.2 DISCUSSION

Tables 2,3, 4, 5 collectively validate both the tightness of the proposed theoretical bounds and the practical observability of the structured credal diameter. Table 2 shows that the general distance bounds in Theorem 4.3 achieve 100% empirical coverage across all evaluated distribution pairs, with zero gaps in pure covariate or pure label shift regimes and moderate slack only in joint-shift settings, confirming correctness and robustness of the decomposition While Table 3 demonstrates that, in the deterministic hard-label regime, the empirical estimator of the diameter is essentially unbiased, with mean absolute errors on the order of 10^-2 and empirical and analytic diameters matching up to three decimal places, indicating fast concentration. Table 4 extends this result to soft-label settings, showing similarly small estimation gaps and negligible signed bias across different probit sharpness levels, thereby confirming the accuracy of the L1-based characterization of the diameter. Finally, Table 5 verifies the noisy-annotator model, where empirical diameters closely track the closed-form theoretical values across increasing noise budgets, with low RMSE and controlled slack, demonstrating robustness of the framework under aleatoric label noise.

Table 3: Finite-sample validation of the exact diameter characterization in Equation (4) for deterministic labeling mechanisms. Across 500 Gaussian covariate distributions and 5 fixed threshold classifiers, the empirical diameter is estimated from $n = 1000$ samples per environment, repeated over 100 runs. Reported metrics summarize the gap between empirical and analytic diameters; confidence intervals are bootstrap-based. Errors are small, symmetric, and centered at zero, confirming unbiasedness and rapid concentration of the empirical estimator.

Metric	Estimate	95% CI
Mean absolute gap	0.0117	[0.0116, 0.0117]
Median absolute gap	0.0097	[0.0096, 0.0098]
RMSE (gap)	0.0148	[0.0147, 0.0149]
Mean gap (signed)	8.98e-06	[-0.0001, 0.0002]
90% abs. gap quantile	0.0244	[0.0242, 0.0246]
95% abs. gap quantile	0.0293	[0.029, 0.0295]
99% abs. gap quantile	0.0384	[0.038, 0.0389]
Mean analytic diameter	0.465	—
Mean empirical diameter	0.465	[0.4648, 0.4651]

Table 4: Finite-sample validation of the soft-label diameter characterization in Equation (5) under probit labeling mechanisms. We compare three probit sharpness settings $\kappa \in \{1, 2, 3\}$ (with the same Gaussian covariate setup and fixed labelers across runs). Reported metrics summarize the gap between the empirical diameter (estimated from finite samples) and the analytic diameter, including mean/median absolute gap, RMSE, and tail quantiles of the absolute gap; brackets denote 95% bootstrap confidence intervals. Mean signed gaps are near zero, indicating negligible bias.

Metric	$\kappa = 1$ (est. [CI])	$\kappa = 2$ (est. [CI])	$\kappa = 3$ (est. [CI])
Mean absolute gap	0.0084 [0.0084, 0.0085]	0.0101 [0.01, 0.0101]	0.0106 [0.0105, 0.0107]
Median absolute gap	0.007 [0.0069, 0.007]	0.0084 [0.0083, 0.0084]	0.0088 [0.0087, 0.0089]
RMSE (gap)	0.0107 [0.0107, 0.0108]	0.0128 [0.0127, 0.0129]	0.0135 [0.0134, 0.0135]
Mean gap (signed)	3.60e-05 [-7.00e-05, 0.0001]	3.63e-05 [-9.37e-05, 0.0002]	3.65e-05 [-9.91e-05, 0.0002]
90% abs. gap quantile	0.0176 [0.0175, 0.0178]	0.0211 [0.0209, 0.0212]	0.0222 [0.022, 0.0223]
95% abs. gap quantile	0.0212 [0.0211, 0.0214]	0.0253 [0.0251, 0.0255]	0.0265 [0.0263, 0.0267]
99% abs. gap quantile	0.0281 [0.0278, 0.0284]	0.0335 [0.0332, 0.0339]	0.035 [0.0346, 0.0354]
Mean analytic diameter	0.4455 (—)	0.4599 (—)	0.4627 (—)
Mean empirical diameter	0.4456 [0.4455, 0.4457]	0.4599 [0.4598, 0.4601]	0.4627 [0.4626, 0.4629]

Table 5: Noisy-annotator ablation for Eq. (7) across noise budgets $\varepsilon_{\max} \in \{0.1, 0.25, 0.5\}$. Entries report estimate [95% bootstrap CI] for the true diameter, empirical diameter, and error/slack metrics.

Metric	$\varepsilon_{\max} = 0.1$	$\varepsilon_{\max} = 0.25$	$\varepsilon_{\max} = 0.5$
True diameter (closed form)	0.1764 (—)	0.3694 (—)	0.4998 (—)
Mean empirical diameter	0.1783 [0.1782, 0.1783]	0.3716 [0.3715, 0.3717]	0.5212 [0.5211, 0.5212]
Upper bound	0.1772 (—)	0.3706 (—)	0.4998 (—)
Mean gap to true (signed)	0.0018 [0.0017, 0.0019]	0.0022 [0.0021, 0.0023]	0.0214 [0.0213, 0.0215]
Mean abs. gap to true	0.0087 [0.0086, 0.0088]	0.0108 [0.0108, 0.0109]	0.0214 [0.0213, 0.0215]
RMSE to true	0.0109 [0.0109, 0.011]	0.0136 [0.0136, 0.0137]	0.0235 [0.0234, 0.0235]
95% abs. gap quantile (to true)	0.0214 [0.0212, 0.0216]	0.0267 [0.0265, 0.0269]	0.0381 [0.0379, 0.0384]
99% abs. gap quantile (to true)	0.0283 [0.0279, 0.0286]	0.0353 [0.0348, 0.0357]	0.0456 [0.0452, 0.046]
Upper-bound violation rate	0.5162 [0.5123, 0.5202]	0.5167 [0.5125, 0.5208]	0.9935 [0.9927, 0.9942]
Mean slack to upper (signed)	0.001 [0.0009, 0.0011]	0.001 [0.0009, 0.0011]	0.0213 [0.0212, 0.0214]
Mean abs. slack to upper	0.0086 [0.0086, 0.0087]	0.0107 [0.0107, 0.0108]	0.0213 [0.0213, 0.0214]

Tables 6, 7, 8, and 9 empirically validate the predicted sample complexity behavior of the diameter estimator in the deterministic hard-label regime across different covariate distributions. As the sample size increases, the median and high-quantile absolute estimation errors consistently decrease at a rate close to $O(n^{-1/2})$, confirming the theoretical concentration guarantees. Across all settings, the empirical violation probability remains effectively zero, indicating that the Hoeffding-based bounds are conservative yet reliable. Moreover, the tightness ratios remain stable and bounded (approximately between 1.7 and 2.5), demonstrating that the proposed concentration bounds are not only asymptotically correct but also practically informative across varying population diameters and covariate shifts.

Table 6: Sample complexity concentration bounds in the deterministic hard label regime under covariate distribution N(0,1). Population diameter $\eta^*_{Y|X} = 0.683$. Repetitions $R = 2000$. Here, n denotes sample size. $\widehat{p}_{\text{viol}}$ is the estimated probability of violations. Let $\epsilon := \left|\eta^*_{Y|X} - \hat{\eta}_{Y|X}\right|$ and $q_\alpha(\epsilon)$ denote the empirical α-quantile of ϵ.

n	$q_{0.5}(\epsilon)$	$q_{0.95}(\epsilon)$	$\epsilon^{\cup}_{\text{Hoeff}}$	$\widehat{p}_{\text{viol}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{low}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{high}}$	Tightness Ratio
10	0.117	0.283	0.547	0.000	0.000	0.002	1.936
30	0.051	0.183	0.316	0.000	0.000	0.002	1.730
100	0.033	0.093	0.173	0.000	0.000	0.002	1.867
500	0.013	0.039	0.077	0.000	0.000	0.002	1.969
1000	0.010	0.030	0.055	0.000	0.000	0.002	1.844
2000	0.007	0.020	0.039	0.000	0.000	0.002	1.954
5000	0.004	0.013	0.024	0.000	0.000	0.002	1.870
10000	0.003	0.009	0.017	0.000	0.000	0.002	1.843
20000	0.002	0.006	0.012	0.000	0.000	0.002	1.970
100000	0.001	0.003	0.005	0.000	0.000	0.002	1.855

Table 7: Sample complexity concentration bounds in the deterministic hard label regime under covariate distribution N(0,2). Population diameter $\eta^*_{Y|X} = 0.383$. Repetitions $R = 2000$. Here, n denotes sample size. $\widehat{p}_{\text{viol}}$ is the estimated probability of violations. Let $\epsilon := \left|\eta^*_{Y|X} - \hat{\eta}_{Y|X}\right|$ and $q_\alpha(\epsilon)$ denote the empirical α-quantile of ϵ.

n	$q_{0.5}(\epsilon)$	$q_{0.95}(\epsilon)$	$\epsilon^{\cup}_{\text{Hoeff}}$	$\widehat{p}_{\text{viol}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{low}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{high}}$	Tightness Ratio
10	0.117	0.283	0.547	0.000	0.000	0.002	1.935
30	0.050	0.183	0.316	0.000	0.000	0.002	1.727
100	0.033	0.097	0.173	0.001	0.000	0.003	1.783
500	0.015	0.043	0.077	0.000	0.000	0.002	1.803
1000	0.011	0.030	0.055	0.001	0.000	0.004	1.820
2000	0.007	0.021	0.039	0.000	0.000	0.002	1.836
5000	0.005	0.013	0.024	0.000	0.000	0.002	1.816
10000	0.003	0.009	0.017	0.000	0.000	0.002	1.836
20000	0.002	0.007	0.012	0.001	0.000	0.003	1.861
100000	0.001	0.003	0.005	0.001	0.000	0.003	1.719

Table 10 obtains empirical concentration behavior in the soft-label regime with a Probit link under $N(0, 1)$ covariates. The median and 95th-percentile estimation errors decrease monotonically with sample size, exhibiting the predicted $O(n^{-1/2})$ scaling. The Hoeffding-based bound remains conservative with zero observed violations across all trials, while the tightness ratio stays stable, indicating a consistent gap between worst-case guarantees and empirical performance.

Figure 3 outlines the empirical concentration behavior of the proposed estimators under different labeling regimes. In Fig. 3a, corresponding to deterministic hard labels, the absolute error decays approximately at the theoretical rate $O(n^{-1/2})$ across all Gaussian input distributions, confirming that the estimator exhibits standard parametric concentration while remaining substantially below the Hoeffding union bound. Fig. 3b highlights that soft labeling mechanisms, including sigmoid and probit models, preserve the same asymptotic scaling; however, the sigmoid mechanism consistently achieves lower error, suggesting improved variance properties due to smoother probabilistic supervision. Finally, Fig. 3c demonstrates the impact

Table 8: Sample complexity concentration bounds in the deterministic hard label regime under covariate distribution N(2,1). Population diameter $\eta^*_{Y|X} = 0.157$. Repetitions $R = 2000$. Here, n denotes sample size. $\widehat{p}_{\text{viol}}$ is the estimated probability of violations. Let $\epsilon := \left|\eta^*_{Y|X} - \hat{\eta}_{Y|X}\right|$ and $q_\alpha(\epsilon)$ denote the empirical α-quantile of ϵ.

n	$q_{0.5}(\epsilon)$	$q_{0.95}(\epsilon)$	$\epsilon^{\cup}_{\text{Hoeff}}$	$\widehat{p}_{\text{viol}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{low}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{high}}$	Tightness Ratio
10	0.057	0.243	0.547	0.000	0.000	0.002	2.255
30	0.043	0.143	0.316	0.000	0.000	0.002	2.215
100	0.027	0.077	0.173	0.000	0.000	0.002	2.239
500	0.011	0.031	0.077	0.000	0.000	0.002	2.473
1000	0.008	0.023	0.055	0.000	0.000	0.002	2.349
2000	0.005	0.016	0.039	0.000	0.000	0.002	2.449
5000	0.004	0.010	0.024	0.000	0.000	0.002	2.332
10000	0.002	0.007	0.017	0.000	0.000	0.002	2.402
20000	0.002	0.005	0.012	0.000	0.000	0.002	2.474
100000	0.001	0.002	0.005	0.000	0.000	0.002	2.406

Table 9: Labeling mechanism complexity concentration bounds in the deterministic hard label regime (Interval Method) under covariate distribution N(2,1). Fixed sample size $n = 1000$. Repetitions $R = 2000$. Here, N_Y denotes the number of labeling mechanisms. $\widehat{p}_{\text{viol}}$ is the estimated probability of violations. Let $\epsilon := \left|\eta^*_{Y|X} - \hat{\eta}_{Y|X}\right|$ and $q_\alpha(\epsilon)$ denote the empirical α-quantile of ϵ. The union Hoeffding bound is $\epsilon^{\cup}_{\text{Hoeff}}$.

N_Y	$q_{0.5}(\epsilon)$	$q_{0.95}(\epsilon)$	$\epsilon^{\cup}_{\text{Hoeff}}$	$\widehat{p}_{\text{viol}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{low}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{high}}$	Tightness Ratio
2	0.003	0.009	0.043	0.000	0.000	0.002	4.605
5	0.007	0.022	0.055	0.000	0.000	0.002	2.496
12	0.008	0.022	0.063	0.000	0.000	0.002	2.820
20	0.008	0.022	0.067	0.000	0.000	0.002	3.004
30	0.008	0.022	0.070	0.000	0.000	0.002	3.226
50	0.008	0.022	0.073	0.000	0.000	0.002	3.336
80	0.008	0.022	0.077	0.000	0.000	0.002	3.409
100	0.007	0.023	0.078	0.000	0.000	0.002	3.328
200	0.007	0.022	0.082	0.000	0.000	0.002	3.825
500	0.008	0.022	0.088	0.000	0.000	0.002	3.986
1000	0.008	0.023	0.092	0.000	0.000	0.002	3.991

Table 10: Sample complexity concentration bounds in the soft-label regime using the Probit link function under the covariate distribution N(0,1). Population diameter $\eta^*_{Y|X} = 0.520$. Repetitions $R = 100$. Here, n denotes sample size. $\widehat{p}_{\text{viol}}$ is the estimated probability of violations. Let $\epsilon := \left|\eta^*_{Y|X} - \hat{\eta}_{Y|X}\right|$ and $q_\alpha(\epsilon)$ denote the empirical α-quantile of ϵ.

n	$q_{0.5}(\epsilon)$	$q_{0.95}(\epsilon)$	$\epsilon^{\cup}_{\text{Hoeff}}$	$\widehat{p}_{\text{viol}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{low}}$	$\text{CI}^{\text{Wilson}}_{0.95,\text{high}}$	Tightness Ratio
10	0.038	0.104	0.547	0.000	0.000	0.037	5.269
30	0.022	0.057	0.316	0.000	0.000	0.037	5.514
100	0.011	0.032	0.173	0.000	0.000	0.037	5.420
500	0.005	0.012	0.077	0.000	0.000	0.037	6.399
1000	0.003	0.009	0.055	0.000	0.000	0.037	6.034
2000	0.002	0.006	0.039	0.000	0.000	0.037	6.070
10000	0.001	0.003	0.017	0.000	0.000	0.037	6.162
20000	0.001	0.002	0.012	0.000	0.000	0.037	6.289
100000	0.000	0.001	0.005	0.000	0.000	0.037	5.559

of deterministic labeling structure: interval-based mechanisms maintain uniformly low error as the sample size increases, whereas block-based labeling converges more slowly and exhibits a higher error floor. Together, our results indicate that while the asymptotic rate is robust to the choice of labeling strategy, the constant factors and finite-sample behavior are strongly influenced by the labeling mechanism and data distribution.

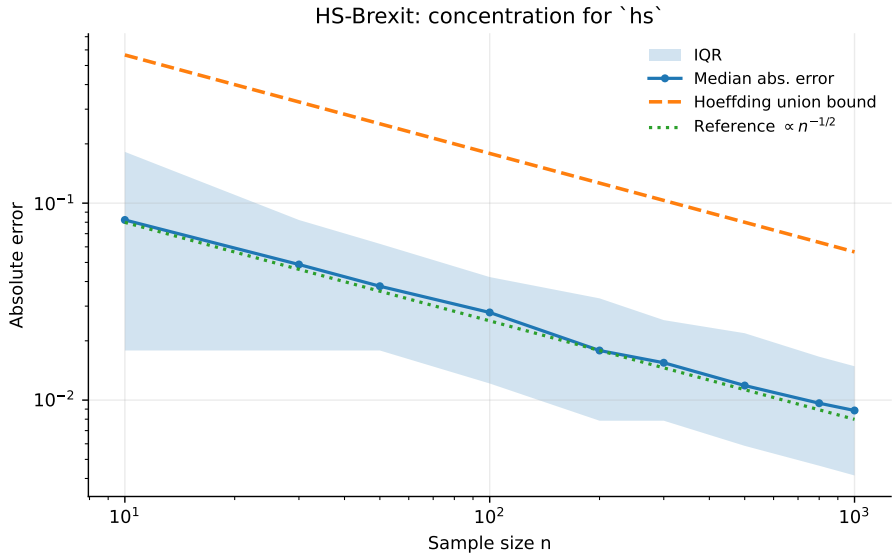
E EMPIRICAL STUDY ON REAL-WORLD DATASET

Across all HS-Brexit annotation dimensions, the empirical disagreement-diameter estimator exhibits stable finite-sample behavior consistent with Theorem 5.2. For hard annotator labels, the empirical population diameters are 0.218 for hate speech, 0.188 for aggressiveness, 0.302 for offensiveness, and 0.265 for stereotypes. In all cases, the median absolute error decreases approximately at the predicted $O(n^{-1/2})$ rate, reaching about 0.008–0.010 at $n = 1000$. The empirical 95th-percentile errors remain below the Hoeffding union-bound envelope, with essentially no observed violations across 2000 repetitions.

E.1 LABEL: `HS` (HATE SPEECH)

Table 11: Real-world concentration of the empirical annotator-disagreement diameter on HS-Brexit. Results are shown for the `hs` label. The empirical population contains $N = 1120$ tweet instances. There are $K = 6$ annotators. The empirical population diameter is $\eta^*_{Y|X} = 0.218$. The worst annotator pair in the empirical population is `annotator_2-annotator_5`. Each row uses $R = 2000$ i.i.d. repetitions. For each n, samples are drawn with replacement from the empirical distribution over tweet instances. Here, $\epsilon = |\eta^*_{Y|X} - \widehat{\eta}_{Y|X}|$. $q_\alpha(\epsilon)$ denotes the empirical α-quantile of the absolute error. $\epsilon_n(\delta)$ is the Hoeffding union-bound envelope. $\widehat{p}_{\text{viol}}$ is the empirical violation probability.

n	$\widehat{\eta}^{\text{med}}_{Y\|X}$	$q_{.50}(\epsilon)$	$q_{.95}(\epsilon)$	$\epsilon_n(\delta)$	$\widehat{p}_{\text{viol}}$	CI^-_{95}	CI^+_{95}	$\epsilon_n/q_{.95}(\epsilon)$
10	0.300	0.082	0.282	0.566	0.000	0.000	0.002	2.004
30	0.233	0.049	0.149	0.327	0.000	0.000	0.002	2.194
50	0.240	0.038	0.122	0.253	0.000	0.000	0.002	2.071
100	0.230	0.028	0.082	0.179	0.000	0.000	0.002	2.177
200	0.225	0.018	0.057	0.126	0.000	0.000	0.002	2.213
300	0.220	0.015	0.045	0.103	0.000	0.000	0.002	2.271
500	0.220	0.012	0.036	0.080	0.000	0.000	0.002	2.213
800	0.220	0.010	0.028	0.063	0.000	0.000	0.002	2.227
1000	0.219	0.009	0.025	0.057	0.000	0.000	0.002	2.249



E.2 LABEL: `AGRESSIVENESS`

Table 12: Real-world concentration of the empirical annotator-disagreement diameter on HS-Brexit. Results are shown for the `aggressiveness` label. The empirical population contains $N = 1120$ tweet instances. There are $K = 6$ annotators. The empirical population diameter is $\eta^*_{Y|X} = 0.188$. The worst annotator pair in the empirical population is `annotator_3-annotator_5`. Each row uses $R = 2000$ i.i.d. repetitions. For each n, samples are drawn with replacement from the empirical distribution over tweet instances. Here, $\epsilon = |\eta^*_{Y|X} - \widehat{\eta}_{Y|X}|$. $q_\alpha(\epsilon)$ denotes the empirical α-quantile of the absolute error. $\epsilon_n(\delta)$ is the Hoeffding union-bound envelope. $\widehat{p}_{\text{viol}}$ is the empirical violation probability.

n	$\widehat{\eta}^{\text{med}}_{Y\|X}$	$q_{.50}(\epsilon)$	$q_{.95}(\epsilon)$	$\epsilon_n(\delta)$	$\widehat{p}_{\text{viol}}$	CI^-_{95}	CI^+_{95}	$\epsilon_n / q_{.95}(\epsilon)$
10	0.300	0.112	0.312	0.566	0.000	0.000	0.002	1.815
30	0.233	0.045	0.145	0.327	0.000	0.000	0.002	2.253
50	0.220	0.032	0.112	0.253	0.000	0.000	0.002	2.266
100	0.200	0.022	0.078	0.179	0.000	0.000	0.002	2.281
200	0.195	0.017	0.052	0.126	0.000	0.000	0.002	2.450
300	0.193	0.015	0.042	0.103	0.000	0.000	0.002	2.482
500	0.192	0.012	0.034	0.080	0.000	0.000	0.002	2.380
800	0.190	0.008	0.025	0.063	0.000	0.000	0.002	2.494
1000	0.190	0.008	0.024	0.057	0.000	0.000	0.002	2.396

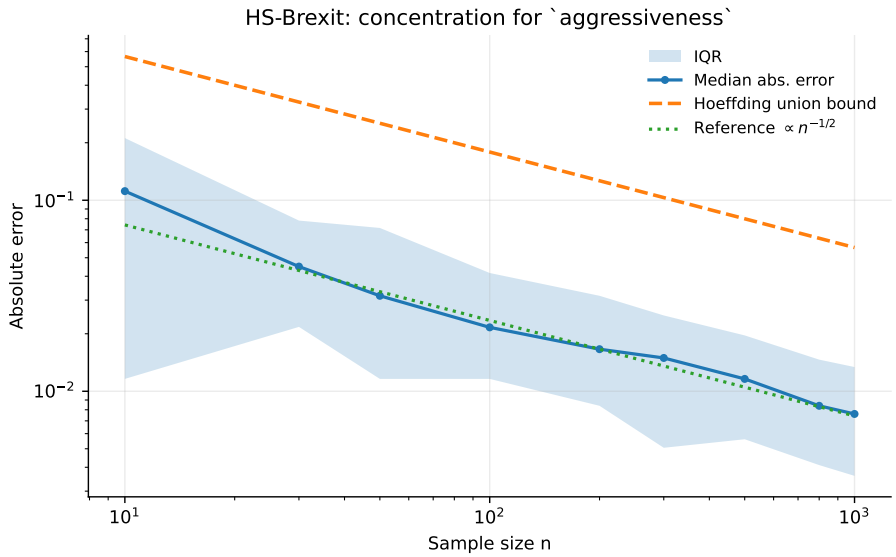


Table 13: Real-world concentration of the empirical annotator-disagreement diameter on HS-Brexit. Results are shown for the `offensiveness` label. The empirical population contains $N = 1120$ tweet instances. There are $K = 6$ annotators. The empirical population diameter is $\eta^*_{Y|X} = 0.302$. The worst annotator pair in the empirical population is `annotator_3-annotator_5`. Each row uses $R = 2000$ i.i.d. repetitions. For each n, samples are drawn with replacement from the empirical distribution over tweet instances. Here, $\epsilon = |\eta^*_{Y|X} - \widehat{\eta}_{Y|X}|$. $q_\alpha(\epsilon)$ denotes the empirical α-quantile of the absolute error. $\epsilon_n(\delta)$ is the Hoeffding union-bound envelope. $\widehat{p}_{\text{viol}}$ is the empirical violation probability.

n	$\widehat{\eta}^{\text{med}}_{Y\|X}$	$q_{.50}(\epsilon)$	$q_{.95}(\epsilon)$	$\epsilon_n(\delta)$	$\widehat{p}_{\text{viol}}$	CI^-_{95}	CI^+_{95}	$\epsilon_n / q_{.95}(\epsilon)$
10	0.400	0.098	0.298	0.566	0.000	0.000	0.002	1.896
30	0.367	0.065	0.165	0.327	0.000	0.000	0.002	1.980
50	0.340	0.042	0.118	0.253	0.001	0.000	0.003	2.140
100	0.320	0.032	0.088	0.179	0.000	0.000	0.002	2.027
200	0.310	0.022	0.062	0.126	0.000	0.000	0.002	2.047
300	0.307	0.018	0.045	0.103	0.000	0.000	0.002	2.288
500	0.304	0.014	0.038	0.080	0.000	0.000	0.002	2.093
800	0.302	0.011	0.031	0.063	0.000	0.000	0.002	2.059
1000	0.302	0.010	0.028	0.057	0.000	0.000	0.002	2.034

E.3 LABEL: `OFFENSIVENESS`

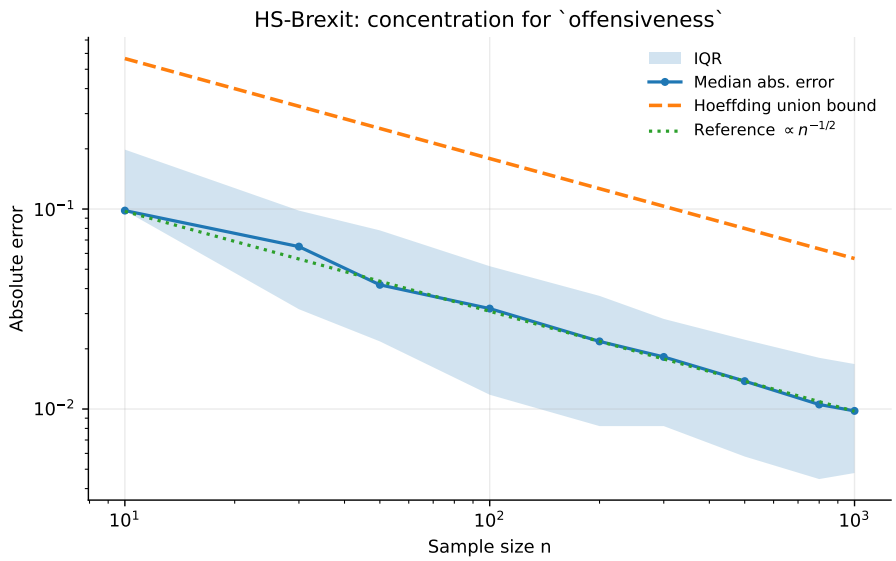
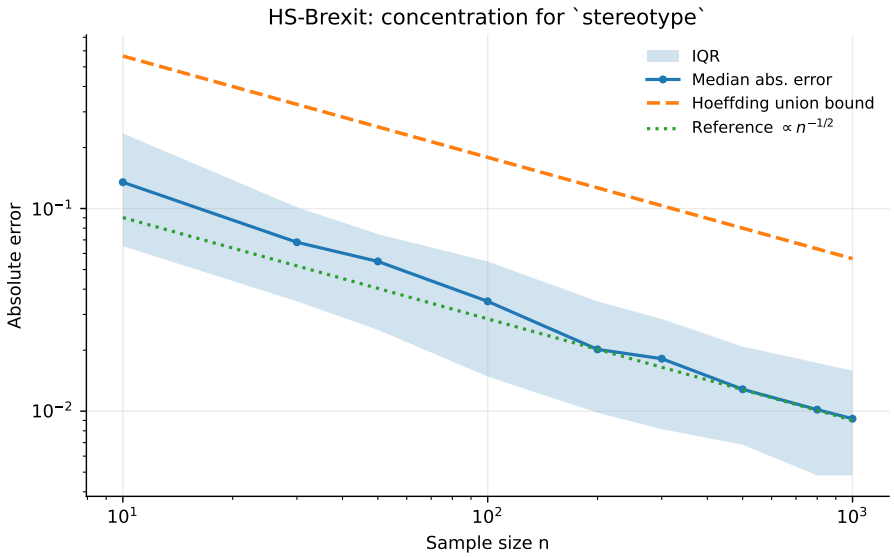
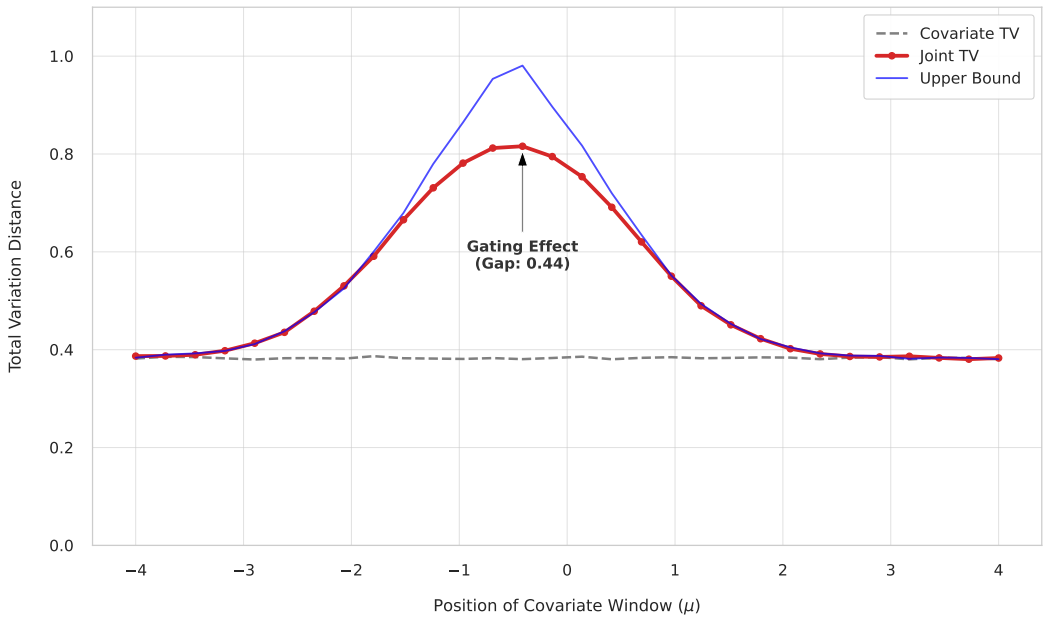


Table 14: Real-world concentration of the empirical annotator-disagreement diameter on HS-Brexit. Results are shown for the `stereotype` label. The empirical population contains $N = 1120$ tweet instances. There are $K = 6$ annotators. The empirical population diameter is $\eta^*_{Y|X} = 0.265$. The worst annotator pair in the empirical population is `annotator_3-annotator_6`. Each row uses $R = 2000$ i.i.d. repetitions. For each n, samples are drawn with replacement from the empirical distribution over tweet instances. Here, $\epsilon = |\eta^*_{Y|X} - \widehat{\eta}_{Y|X}|$. $q_\alpha(\epsilon)$ denotes the empirical α-quantile of the absolute error. $\epsilon_n(\delta)$ is the Hoeffding union-bound envelope. $\widehat{p}_{\text{viol}}$ is the empirical violation probability.

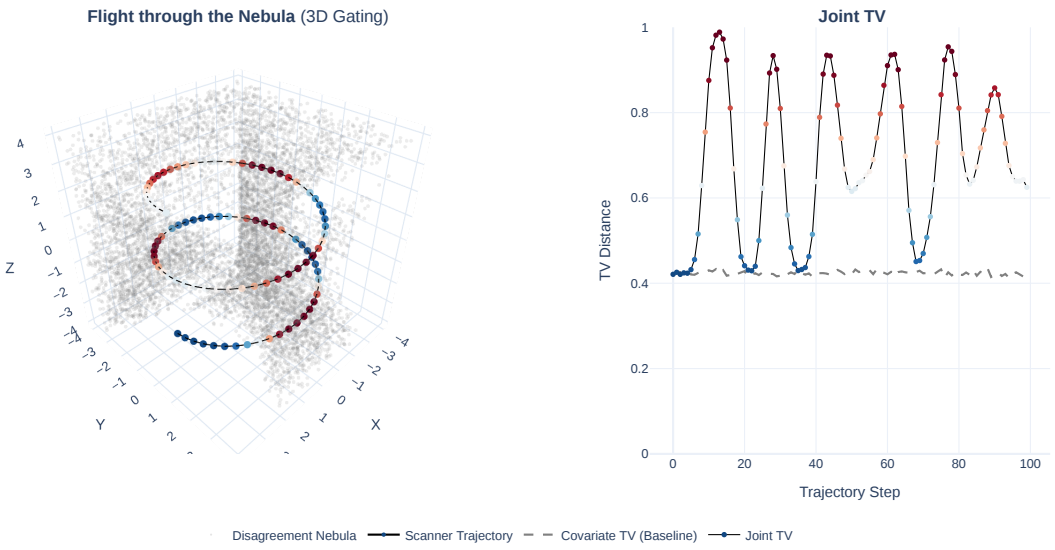
n	$\widehat{\eta}^{\text{med}}_{Y\|X}$	$q_{.50}(\epsilon)$	$q_{.95}(\epsilon)$	$\epsilon_n(\delta)$	$\widehat{p}_{\text{viol}}$	CI^-_{95}	CI^+_{95}	$\epsilon_n/q_{.95}(\epsilon)$
10	0.400	0.135	0.335	0.566	0.001	0.000	0.003	1.689
30	0.333	0.068	0.168	0.327	0.000	0.000	0.002	1.942
50	0.300	0.055	0.135	0.253	0.001	0.000	0.004	1.876
100	0.290	0.035	0.095	0.179	0.000	0.000	0.002	1.886
200	0.280	0.020	0.065	0.126	0.000	0.000	0.002	1.951
300	0.277	0.018	0.051	0.103	0.000	0.000	0.002	2.005
500	0.273	0.013	0.039	0.080	0.000	0.000	0.002	2.060
800	0.270	0.010	0.030	0.063	0.000	0.000	0.002	2.120
1000	0.270	0.009	0.028	0.057	0.000	0.000	0.002	2.033

E.4 LABEL: `STEREOTYPE`



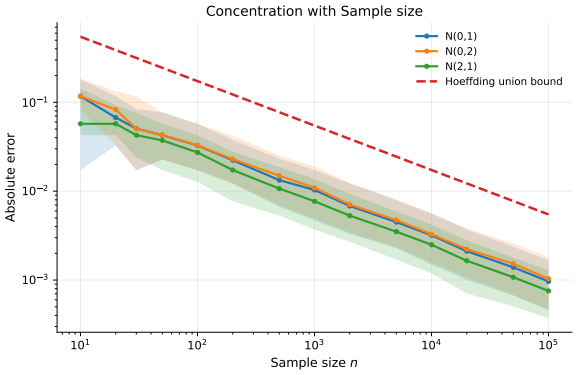


(a) **Gating Effect under Covariate Shift.** Joint total variation distance varies with the position of the covariate window, peaking when probability mass overlaps regions of high labeling disagreement. While the covariate-only TV remains nearly constant, the joint TV closely follows the theoretical upper bound, illustrating how covariate shift modulates the observable impact of labeling uncertainty.

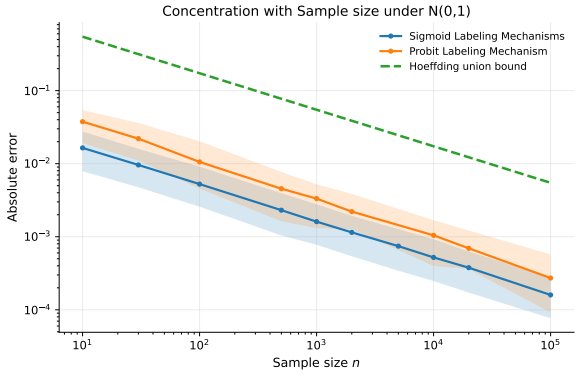


(b) **3D Gating Visualization**: The left panel illustrates a scanner trajectory passing through a three-dimensional disagreement region, while the right panel shows the corresponding joint total variation (TV) distance along the trajectory. Peaks in joint TV occur when the trajectory intersects high-disagreement regions, whereas outside these regions the divergence remains close to the covariate-only baseline, demonstrating the gating effect of covariate mass on observable labeling uncertainty.

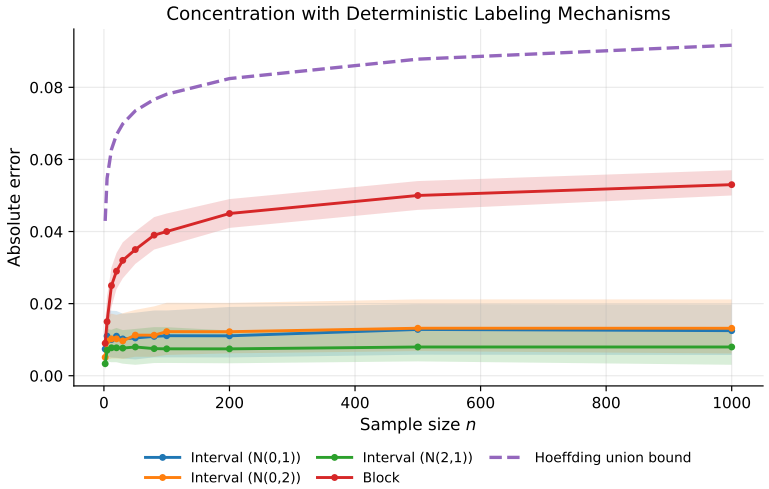
Figure 2: Proposition 4.3 shows that joint distributional divergence results from the interaction between covariate shift and labeling disagreement, with bounds corresponding to two interpolation paths between feature and label changes. Label disagreement affects the joint distance only on regions with non-negligible covariate mass, yielding a **gating effect** in which covariate shift controls the visibility of labeling uncertainty. Consequently, training distributions may underestimate risk when deployment concentrates probability mass in high-disagreement regions, a coupling made explicit by the structured credal framework.



(a) **Deterministic hard labels.** Absolute estimation error versus sample size (log–log scale) for Gaussian inputs with varying variances. The curves demonstrate the canonical $O(n^{-1/2})$ concentration rate, with empirical errors lying well below the Hoeffding union bound.



(b) **Soft (probabilistic) label.** Absolute estimation error versus sample size (log–log scale) under sigmoid and probit labeling mechanisms for $N(0, 1)$ inputs. Both mechanisms exhibit $O(n^{-1/2})$ convergence, with sigmoid labeling achieving systematically lower error.



(c) **Deterministic labeling mechanisms.** Absolute error as a function of sample size (linear scale) for interval-based and block-based deterministic labelers. Interval mechanisms yield stable, low error, while block labeling shows slower improvement and larger asymptotic error; the Hoeffding union bound provides a conservative upper envelope.

Figure 3: Empirical concentration behavior of estimation error under different data distributions and labeling mechanisms. Across settings, the observed error exhibits the expected theoretical scaling with sample size, and is compared against Hoeffding-type union bounds to highlight the gap between empirical performance and worst-case guarantees.