Information-Theoretic Lower Bounds for Causal Inference under Credal Uncertainty

Hung Mai[1,2] **Hai Nguyen**[3] **Khanh Nguyen**[*1,3] **Luong Doan**[1] **Nhung Duong**[1] **Phong Ho**[1] **Tuan Do**[1,3]

[1]N2TP Technology Solutions JSC, Vietnam
[2]National Economics University, Vietnam
[3]Phenikaa University, Vietnam

Abstract

Causal effect estimation from observational data typically assumes a single fixed observational distribution. We study the setting in which the distribution is known only to belong to a credal set, a convex set of plausible observational laws. Our results characterize how this observational ambiguity is transformed by causal identification formulas. First, using standard two-point information-theoretic tools after composition with the interventional map, we derive lower bounds for causal effect estimation. The key message is a causal amplification phenomenon: observational discrepancies in low-propensity treatment strata can be hard to detect while producing separated interventional effects. Consequently, in the binary hard-pair constructions, sample complexity scales as $\Omega(1/(\pi_0\varepsilon^2))$ or $\Omega(1/(\beta_0\varepsilon^2))$, and the average treatment effect need not incur an additional dependence on the number of strata. Second, we prove Lipschitz bounds for the interventional mapping, with amplification factor $(1 + 1/\alpha_{\min})$, showing how positivity controls robust identification width. Third, we study minimax regret for causal decisions under credal uncertainty and show that randomization can halve worst-case regret in a symmetric action-separation construction. We also give matching upper bounds in special cases and an NP-hardness result for computing robust identification width. The experiments are diagnostic checks of the tight hard-instance scaling laws rather than broad empirical benchmarks.

Keywords: credal sets, causal inference, sample complexity, information-theoretic lower bounds, minimax regret, imprecise probabilities

*Correspondence: khanh.nguyenquoc@phenikaa-uni.edu.vn

1 INTRODUCTION

Causal inference asks what would happen to an outcome Y if we intervened and set a treatment X to a value x. Under standard structural assumptions, the interventional distribution can be identified from a single observational distribution P by formulas such as adjustment [Pearl, 2000]. In many applications, however, the observational law is not known exactly: it is estimated from finite samples, may vary across domains, or may only be partially constrained by substantive knowledge. We model this ambiguity by a credal set $\mathcal{P}$, a convex set of plausible observational distributions [Walley, 1991, Augustin et al., 2014].

In causal workflows, credal sets may arise from several sources: finite-sample confidence regions around an estimated observational distribution; multi-domain estimation where the target population is assumed to lie in the convex hull or neighborhood of observed domains; covariate-shift uncertainty over $P(\mathbf{Z})$; interval or moment constraints on mechanisms such as $P(X \mid \mathbf{Z})$ or $P(Y \mid X, \mathbf{Z})$; missing-data constraints; and sensitivity-analysis constraints. We take $\mathcal{P}$ as given and study the information-theoretic consequences of this uncertainty. Learning or estimating $\mathcal{P}$ from data is an important complementary problem, but it is not the focus of this paper.

A first intuition is that larger credal sets should make causal estimation harder. Our results show that this intuition is incomplete. The diameter of $\mathcal{P}$ is a feasibility condition: to prove a nontrivial lower bound, $\mathcal{P}$ must contain distributions with separated causal effects. The statistical rate, however, is governed by how the causal map weights observational discrepancies. In particular, low treatment propensity can attenuate the observational divergence between two worlds while leaving their interventional distributions separated. Thus positivity, not only credal diameter, controls the difficulty of causal estimation and decision-making under credal uncertainty.

The technical ingredients used for lower bounds, such as

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4252–4271.

Le Cam's method, KL inequalities, and elementary total-variation facts, are standard. The contribution is to apply these tools after composition with the causal interventional map and to make explicit the resulting causal amplification factors. This organization also clarifies which results are statistical templates and which statements are causal consequences of identification through adjustment.

We make three main contributions. First, we prove two-point lower bounds for causal effect estimation under credal uncertainty. Proposition 2 is a warm-up no-confounding binary construction; Corollary 5 states the general adjustment lower-bound template for arbitrary hard pairs; and Corollary 6 gives an explicit least-favorable construction where the bound scales as $\Omega(1/(\beta_0\varepsilon^2)\log(1/\delta))$. Matching plug-in and IPW upper bounds are given in Theorems 26 and 27. Second, we quantify how observational uncertainty propagates to interventional uncertainty via the robust identification width $W(\mathcal{P}, x)$, proving the Lipschitz bound in Theorem 15 and showing that computing W is NP-hard in Theorem 19. Third, we study causal decision-making under credal ambiguity: Theorem 23 gives a pure-action lower bound and Theorem 24 shows that randomization achieves a factor-two improvement in a symmetric least-favorable construction.

2 RELATED WORK

Causal identifiability under a fixed distribution. Classical causal inference asks when interventional quantities are identifiable from a single observational distribution. Foundational results by Pearl [2000], Shpitser and Pearl [2006], and Bareinboim and Pearl [2016] characterize graphical conditions for identifying $P(Y \mid \mathrm{do}(X = x))$, with extensions to partial observability, constrained model classes, selection bias, and broader structural assumptions [Zhang et al., 2016, Lee and Bareinboim, 2020, Bodik and Chavez-Demoulin, 2025]. Our work instead allows the observational distribution to belong to a credal set and studies estimation and decision-making when only a set of compatible distributions is known.

Imprecise probabilities and credal models. Credal sets and imprecise probability theory formalize partially specified distributions [Walley, 1991, Augustin et al., 2014]. Credal networks [Cozman, 2000] extend Bayesian networks to sets of distributions, while minimax testing under capacities [Huber and Strassen, 1973] gives early principles for inference under ambiguity. Recent work studies credal semantics in probabilistic logic programming, higher-order credal models, uncertain reasoning, and argumentation [Azzolini and Riguzzi, 2024, Morveli-Espinoza et al., 2019, Hibshman and Weninger, 2021]. We integrate these ideas with causal inference by analyzing how credal uncertainty propagates through interventional maps.

Learning and robustness under distributional uncertainty. Credal learning theory [Caprio et al., 2024] gives generalization bounds when training data are drawn from a credal set, and distributionally robust optimization [Duchi and Namkoong, 2021] studies optimization over uncertainty sets. Our results connect these perspectives to causal inference through sample-complexity lower bounds, Lipschitz sensitivity bounds, and regret guarantees for causal decisions under credal ambiguity.

Partial identification and sensitivity analysis. Partial identification [Manski, 2003] characterizes sets of plausible parameter values under weak assumptions, while sensitivity analysis [Rosenbaum, 2002, Zhao et al., 2019] studies robustness to unobserved confounding. Modern extensions address measurement error and continuous treatments [Finkelstein et al., 2021, Dalal and Tchetgen, 2025]. Our robust identification width complements this work by quantifying how TV uncertainty in the observational distribution is amplified by intervention through explicit Lipschitz bounds.

Information-theoretic and decision-theoretic perspectives. Information-theoretic tools have been used for causal discovery and graph identifiability [Kocaoglu et al., 2020, Compton et al., 2022]; we use them to derive minimax sample-complexity bounds under credal uncertainty. On the decision side, prior work on imprecise probabilities [Troffaes, 2007, Huntley et al., 2014] studies general decision criteria, whereas we give explicit minimax regret bounds for causal interventions.

3 PRELIMINARIES

We briefly fix notation; standard background is deferred to Appendix A.

Scope and variables. The main non-asymptotic results are stated for finite discrete variables. The treatment X is discrete, and interventions are of the form $\mathrm{do}(X = x)$ for a treatment level x in a finite action set. The sharp hard-pair constructions use binary outcomes $Y \in \{0, 1\}$; extensions to multi-valued or continuous outcomes require additional regularity and are discussed as future work.

Causal estimand. We consider variables $(X, Y, \mathbf{Z})$ generated by an SCM with a valid backdoor adjustment set $\mathbf{Z}$. The causal effect of interest is

$$\theta(P) = \mathbb{E}_P[Y \mid \mathrm{do}(X = x)]$$
$$= \sum_{\mathbf{z}} \mathbb{E}_P[Y \mid X = x, \mathbf{Z} = \mathbf{z}] P(\mathbf{Z} = \mathbf{z}). \quad (1)$$

For binary Y, this is equivalently $\sum_{\mathbf{z}} P(Y = 1 \mid X = x, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z})$.

Credal set. The observational distribution is unknown and assumed to lie in a closed convex credal set $\mathcal{P}$ over $(X, Y, \mathbf{Z})$. Its total-variation diameter is

$$\begin{aligned}\operatorname{diam}_{TV}(\mathcal{P}) &= \sup_{P,Q\in\mathcal{P}} d_{TV}(P,Q), \\ d_{TV}(P,Q) &= \tfrac{1}{2}\sum_{\omega} |P(\omega)-Q(\omega)|.\end{aligned} \tag{2}$$

We use $\mathcal{P}$ as an uncertainty model over observational laws, not as a replacement for causal structure: all causal functionals below are assumed identifiable for each $P \in \mathcal{P}$ under the stated graph or adjustment assumptions.

Information-theoretic quantities. We use KL divergence D_{KL} and the Bretagnolle–Huber inequality:

$$d_{TV}(P^n, Q^n)^2 \leq 1 - \exp(-nD_{KL}(P\|Q)). \tag{3}$$

Minimax estimation. An estimator $\hat{\theta}_n$ is (ε, δ)-correct over $\mathcal{P}$ if

$$\sup_{P^*\in\mathcal{P}} P^{*n}\Big(|\hat{\theta}_n - \theta(P^*)| > \varepsilon\Big) \leq \delta. \tag{4}$$

4 SAMPLE COMPLEXITY LOWER BOUNDS UNDER CREDAL UNCERTAINTY

We first study the minimax sample complexity of estimating an identifiable causal functional when the observational distribution lies in $\mathcal{P}$. The lower-bound logic is standard: if $\mathcal{P}$ contains two observationally close distributions with separated causal effects, no estimator can distinguish them reliably with too few samples. The causal content is in how such pairs are constructed. Low treatment propensity makes differences in $P(Y \mid X = x, \mathbf{Z})$ rare observationally, hence small in KL, while the same differences remain fully visible after the intervention $\operatorname{do}(X = x)$.

4.1 TWO-POINT LOWER BOUND VIA LE CAM'S METHOD

We first record the standard two-point testing step used throughout the lower-bound arguments. The causal content in later results is not the Le Cam inequality itself, but the construction of observationally close worlds whose interventional functionals are separated.

Lemma 1 (Two-Point Lower Bound Template)**.** *Let $\theta : \mathcal{P} \to \mathbb{R}$ be a target functional. Suppose there exist $P_0, P_1 \in \mathcal{P}$ such that $|\theta(P_1) - \theta(P_0)| \geq 2\varepsilon$. Then, for any $\delta \in (0, 1/2)$, every (ε, δ)-correct estimator satisfies*

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{D_{\mathrm{KL}}(P_0\|P_1)}.$$

For $\delta \geq 1/2$, the bound is vacuous.

Proof sketch. By Le Cam's method [LeCam, 1973, Cam, 1986], (ε, δ)-correctness requires $d_{\mathrm{TV}}(P_0^n, P_1^n) \geq 1 - 2\delta$. The Bretagnolle–Huber inequality gives $d_{\mathrm{TV}}(P_0^n, P_1^n) \leq \sqrt{1 - \exp(-nD_{\mathrm{KL}}(P_0\|P_1))}$. Solving for n yields the claim. □

4.2 EXPLICIT BOUND FOR BINARY OUTCOMES

The following proposition is a warm-up construction. In the graph $X \to Y$ without confounding, $P(Y \mid \operatorname{do}(X = x)) = P(Y \mid X = x)$, so the causal functional is a conditional projection of the observational distribution. We keep this result to expose the basic propensity mechanism; the adjustment results below are the main causal lower bounds.

Proposition 2 (Explicit Bound for Binary Outcomes)**.** *Assume $\delta \in (0, \frac{1}{2})$. Consider the case where $\theta(P) = P(Y = 1 \mid X = x)$ with binary Y and no confounding. Suppose $\mathcal{P}$ contains distributions P_0, P_1 such that: (i) $P_0(X = x) = P_1(X = x) = \pi_0$ for some $\pi_0 > 0$; (ii) $P_0(Y = 1 \mid X = x) = p_0$ and $P_1(Y = 1 \mid X = x) = p_1$ with $|p_1 - p_0| \geq 2\varepsilon$; (iii) P_0 and P_1 are identical except on $P(Y \mid X = x)$. Then*

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\pi_0 \cdot D_{KL}(\mathrm{Bern}(p_0)\|\mathrm{Bern}(p_1))}. \tag{5}$$

If additionally $\mathcal{P}$ contains the symmetric hard pair with $p_0 = \frac{1}{2} - \varepsilon$ and $p_1 = \frac{1}{2} + \varepsilon$, then

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\pi_0 \cdot 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}}. \tag{6}$$

For small ε,

$$n = \Omega\left(\frac{1}{\pi_0\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right). \tag{7}$$

Proof sketch. Since P_0 and P_1 are identical except on $P(Y \mid X = x)$, the KL divergence localizes to the treated slice: $D_{KL}(P_0\|P_1) = \pi_0 D_{KL}(\mathrm{Bern}(p_0)\|\mathrm{Bern}(p_1))$. The symmetric-pair expression follows by direct calculation, and Lemma 1 completes the proof. The full derivation appears in Appendix F.2. □

A matching upper bound is established by Theorem 26 in Appendix C.1: the plug-in conditional mean achieves $n = O(1/(\pi_{\min}\varepsilon^2)\log(1/\delta))$, confirming that the lower bound is tight up to constants.

Remark 3 (Least-favorable pairs). The hard pairs above and below are least-favorable submodels, not assumptions that realistic credal sets vary only in one conditional. The lower bound follows by restriction: if estimation is hard on a two-point subset of $\mathcal{P}$, it is also hard on the full credal set. These constructions isolate the causal direction in which observational discrepancies are attenuated by treatment propensity while interventional discrepancies are not.

4.3 ADJUSTMENT SETTINGS: GENERAL TEMPLATE AND EXPLICIT PROPENSITY SCALING

For the adjustment case with $\theta(P) = \sum_z P(Y = 1 \mid X = x, Z = z)P(Z = z)$, the relevant hardness is driven by treatment propensity, not by the joint probability of each stratum alone.

Definition 4 (Positivity Constants for Adjustment)**.** For credal set $\mathcal{P}$ with adjustment set $\mathbf{Z}$, the conditional positivity or propensity is

$$\beta_{\min}(\mathcal{P}) = \inf_{P\in\mathcal{P}} \min_z P(X = x \mid \mathbf{Z} = z), \tag{8}$$

and the effective overlap is

$$\kappa(\mathcal{P}) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\frac{1}{P(X = x \mid \mathbf{Z})}\right]. \tag{9}$$

Corollary 5 (Adjustment lower bound for arbitrary hard pairs)**.** *Let $\theta(P) = \sum_z P(Y = 1 \mid X = x, Z = z)P(Z = z)$ be identifiable by adjustment. Suppose $\mathcal{P}$ contains P_0, P_1 such that $|\theta(P_1) - \theta(P_0)| \geq 2\varepsilon$. If $\delta \in (0, 1/2)$, then any (ε, δ)-correct estimator requires*

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{D_{KL}(P_0\|P_1)}. \tag{10}$$

In particular, whenever a least-favorable pair has observational KL attenuated by the propensity of observing $X = x$, the resulting lower bound inherits the corresponding propensity dependence.

Proof. This is Lemma 1 applied to the adjustment functional θ. The statement is intentionally pairwise: any two observationally close but interventionally separated distributions inside $\mathcal{P}$ certify hardness for the full credal set. □

Corollary 6 (Adjusted case, Propensity-Driven Bound)**.** *Let $\mathcal{G}$ be a causal graph where the effect of X on Y is identifiable via adjustment set $\mathbf{Z}$. Assume $Y \in \{0, 1\}$ and $\delta \in (0, \frac{1}{2})$. Suppose $\mathcal{P}$ contains distributions P_0, P_1 satisfying the uniform-shift construction: (i) $P_0(\mathbf{Z}) = P_1(\mathbf{Z})$; (ii) $P_0(X{=}x \mid \mathbf{Z}{=}z) = P_1(X{=}x \mid \mathbf{Z}{=}z) = e(z)$ for all z, with $\min_z e(z) = \beta_0$; (iii) $P_0(Y{=}1 \mid X{=}x, \mathbf{Z}{=}z) = \frac{1}{2} - \varepsilon$ for all z; (iv) $P_1(Y{=}1 \mid X{=}x, \mathbf{Z}{=}z) = \frac{1}{2} + \varepsilon$ for all z; and (v) all other conditionals are identical. Then*

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{P_0(X{=}x) \cdot 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}}. \tag{11}$$

If additionally $P_0(X{=}x) = \beta_0$, then

$$n = \Omega\left(\frac{1}{\beta_0\varepsilon^2} \log \frac{1}{\delta}\right). \tag{12}$$

Proof sketch. Under the uniform shift construction, $\theta(P_0) = 1/2 - \varepsilon$ and $\theta(P_1) = 1/2 + \varepsilon$. The KL divergence localizes to observations with $X = x$, so $D_{KL}(P_0\|P_1) = P_0(X = x) \cdot 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}$. Applying Corollary 5 yields the bound. The full proof is in Appendix F.3. □

Example 7 (Adjustment hard pair with nonuniform propensity). Let $Z \in \{0, 1\}$ with $P(Z = 0) = P(Z = 1) = 1/2$, and let $P(X = 1 \mid Z = 0) = 0.1$ and $P(X = 1 \mid Z = 1) = 0.3$, so $P(X = 1) = 0.2$. Suppose P_0 and P_1 share $P(Z)$ and $P(X \mid Z)$, and differ only in the treated outcome conditional: $P_0(Y = 1 \mid X = 1, Z = z) = 1/2 - \varepsilon$ and $P_1(Y = 1 \mid X = 1, Z = z) = 1/2 + \varepsilon$ for both $z = 0, 1$. Then, for the adjustment estimand $\theta(P) = \sum_z P(Y = 1 \mid X = 1, Z = z)P(Z = z)$, we have $|\theta(P_1) - \theta(P_0)| = 2\varepsilon$. However, Let $b_\pm = \mathrm{Bern}(1/2 \pm \varepsilon)$. Then

$$\begin{aligned} D_{\mathrm{KL}}(P_0\|P_1) &= P(X = 1)D_{\mathrm{KL}}(b_-\|b_+) \\ &= 0.2\{8\varepsilon^2 + O(\varepsilon^4)\}. \end{aligned}$$

Thus the two worlds are observationally close because the distinguishing conditional is observed only among treated units, while their interventional laws under $do(X = 1)$ remain separated. This illustrates the propensity-dependent amplification behind the lower bound.

The positivity condition is satisfied, for example, by any credal set $\mathcal{P}$ for which there exist constants $c_Z, c_X > 0$ such that, for all $P \in \mathcal{P}$ and all relevant z, $P(Z = z) \geq c_Z$ and $P(X = x \mid Z = z) \geq c_X$. Then

$$\alpha_{\min}(\mathcal{P}) := \inf_{P\in\mathcal{P}} \min_z P(X = x, Z = z) \geq c_Z c_X > 0.$$

Thus $\alpha_{\min}(\mathcal{P}) > 0$ is an overlap condition over the whole credal set, not an automatic property. The critical observation is that the sample complexity for the ATE depends on the marginal treatment probability $P(X = x) = \sum_z e(z)P(z)$, not on the joint positivity $\alpha_{\min} = \min_z P(X = x, Z = z)$. In the canonical equal-mass, constant-propensity design, $\alpha_{\min} = \beta/K$ scales with K, but $P(X{=}x) = \beta$ is independent of K. A matching upper bound via IPW [Horvitz and Thompson, 1952] (Theorem 27, Appendix C.2) confirms this: the IPW estimator achieves $n = O(\kappa(\mathcal{P})\varepsilon^{-2}\log(1/\delta))$, where $\kappa \leq 1/\beta_{\min}$.

This absence of K-scaling is specific to averaged effects. If the goal is to estimate $P(Y \mid X = x, Z = z)$ uniformly for every z, then each stratum must be estimated separately, requiring $\Omega(K)$ samples. The distinction is practically important: ATE estimation averages over strata, while uniform conditional estimation requires adequate samples in every stratum.

Remark 8 (Role of credal diameter). The credal diameter $\Delta = \mathrm{diam}_{TV}(\mathcal{P})$ enters the lower bounds only as a feasibility condition: $\mathcal{P}$ must contain distributions whose causal effects differ by at least 2ε. Conditional on such a hard pair, the sample complexity is governed by observational KL, and the constructions show that this KL can be controlled by positivity/propensity rather than by Δ alone.

5 ROBUST IDENTIFICATION WIDTH UNDER CREDAL UNCERTAINTY

We next characterize how observational uncertainty propagates to interventional uncertainty. With infinitely many samples from a single fixed $P^* \in \mathcal{P}$, an identifiable causal effect is point-identified for that P^*. A different question remains: if the true observational law is only known to lie somewhere in $\mathcal{P}$, how much can the intervention law vary across plausible worlds? This is relevant for multi-domain data, distribution shift, finite-sample uncertainty sets, and partial mechanism knowledge.

5.1 DEFINITIONS

Definition 9 (Interventional Mapping)**.** For a causal graph $\mathcal{G}$ and intervention $\mathrm{do}(X = x)$, define the interventional mapping $\Phi_x : \mathcal{P} \to \Delta(\mathcal{Y})$, $P \mapsto P(Y \mid \mathrm{do}(X = x))$, where $\Delta(\mathcal{Y})$ is the simplex of distributions over outcome Y.

Definition 10 (Robust Identification Width)**.** The robust identification width of the causal effect over credal set $\mathcal{P}$ is

$$W(\mathcal{P}, x) = \sup_{P,Q\in\mathcal{P}} d_{TV}(\Phi_x(P), \Phi_x(Q))\,. \qquad (13)$$

Definition 11 (Robust Identifiability)**.** The causal effect $P(Y \mid \mathrm{do}(X = x))$ is robustly point-identified over $\mathcal{P}$ if $W(\mathcal{P}, x) = 0$, robustly partially identified if $0 < W(\mathcal{P}, x) < 1$, and robustly non-identified if $W(\mathcal{P}, x) = 1$.

The quantity $W(\mathcal{P}, x)$ can exceed $\mathrm{diam}_{TV}(\mathcal{P})$ because adjustment involves conditioning on treatment strata. The threshold for robust non-identification is therefore $W = 1$, the maximum possible TV distance between outcome distributions, not $W = \mathrm{diam}_{TV}(\mathcal{P})$.

5.2 LIPSCHITZ BOUND ON THE INTERVENTIONAL MAPPING

Definition 12 (Joint Positivity)**.** For the Lipschitz analysis, define

$$\alpha_{\min}(\mathcal{P}) = \inf_{P\in\mathcal{P}} \min_{z} P(X = x, \mathbf{Z} = z). \qquad (14)$$

The bounds below assume $\alpha_{\min}(\mathcal{P}) > 0$.

Example 13 (When joint positivity holds). If every $P \in \mathcal{P}$ satisfies $P(\mathbf{Z} = z) \geq \zeta_0$ and $P(X = x \mid \mathbf{Z} = z) \geq \beta_0$ for all relevant cells, then $P(X = x, \mathbf{Z} = z) \geq \zeta_0\beta_0$, hence $\alpha_{\min}(\mathcal{P}) \geq \zeta_0\beta_0$. This condition is an overlap assumption; it is not automatic.

Lemma 14 (Standard TV bound under conditioning)**.** *Let P, Q be distributions on a finite space Ω and let $B \subseteq \Omega$ with $P(B), Q(B) > 0$. Then*

$$d_{\mathrm{TV}}(P(\cdot \mid B), Q(\cdot \mid B)) \leq \frac{d_{\mathrm{TV}}(P, Q)}{\min\{P(B), Q(B)\}}.$$

Proof sketch. This is the standard TV conditioning inequality: restrict both measures to B, normalize, and bound the resulting ℓ_1 difference by the unconditional ℓ_1 difference divided by the smaller conditioning mass. See Appendix F.4. □

Theorem 15 (Lipschitz Bound for the Adjustment Map)**.** *Let the effect of X on Y be identifiable by a finite adjustment set* **Z***, and let $\Phi_x(P)$ denote the interventional law obtained from the adjustment formula. Assume joint positivity $\alpha_{\min}(\mathcal{P}) > 0$. Then, for any $P, Q \in \mathcal{P}$,*

$$d_{\mathrm{TV}}(\Phi_x(P), \Phi_x(Q)) \leq \left(1 + \frac{1}{\alpha_{\min}(\mathcal{P})}\right) d_{\mathrm{TV}}(P, Q).$$

Consequently,

$$W(\mathcal{P}, x) \leq \min\left\{1, \left(1 + \frac{1}{\alpha_{\min}(\mathcal{P})}\right) \mathrm{diam}_{\mathrm{TV}}(\mathcal{P})\right\}.$$

Proof sketch. Apply the adjustment formula and split the difference into the change in adjustment weights and the change in conditionals $P(Y \mid X = x, \mathbf{Z} = z)$. The former is bounded by $d_{\mathrm{TV}}(P, Q)$, while the latter is bounded by Lemma 14 and positivity by $d_{\mathrm{TV}}(P, Q)/\alpha_{\min}(\mathcal{P})$. This yields the Lipschitz bound; taking the supremum over $P, Q \in \mathcal{P}$ gives the width bound. □

Proof sketch. By adjustment, $\Phi_x(P) = \sum_z P(\cdot \mid X = x, Z = z)P(Z = z)$. The TV distance between the two mixtures is bounded by the difference in stratum weights plus the weighted difference between conditionals. The stratum-weight term is at most $d_{TV}(P, Q)$, while each conditional term is bounded by Lemma 14 and joint positivity. The full proof is in Appendix F.5. □

Theorem 15 is the causal robustness statement; Lemma 14 is only the standard TV ingredient. The theorem shows that the adjustment map can amplify observational uncertainty by a factor controlled by joint positivity. The bound is worst-case: for benign credal sets, such as small balls around well-conditioned distributions, the empirical width may be far below the Lipschitz upper bound.

Corollary 16 (Bounded Robust Width under Strong Positivity)**.** *If $\alpha_{\min}(\mathcal{P}) \geq \alpha_0 > 0$, then $W(\mathcal{P}, x) \leq \min\{1, (1 + 1/\alpha_0)\mathrm{diam}_{TV}(\mathcal{P})\}$. If the credal diameter Δ is small compared to $\alpha_0/(1 + \alpha_0)$, the robust identification width is also small.*

Corollary 17 (Near-Positivity Allows Maximal Width). *For every $\eta > 0$, there exists a closed, convex credal set $\mathcal{P}_\eta$ such that $\operatorname{diam}_{TV}(\mathcal{P}_\eta) \leq \eta$ and $\alpha_{\min}(\mathcal{P}_\eta) = \eta$, yet $W(\mathcal{P}_\eta, x) = 1$. Consequently, no bound of the form $W(\mathcal{P}, x) \leq C \operatorname{diam}_{TV}(\mathcal{P})$ can hold uniformly over classes with arbitrarily small $\alpha_{\min}$, and the factor $1/\alpha_{\min}$ in Theorem 15 is unavoidable.*

Proof. Take Z degenerate, $X \in \{0,1\}$ with $P(X = 1) = \eta$, and $Y \in \{0,1\}$. Let P_0 and P_1 agree except on $\{X = 1\}$: set $P_0(Y = 1 \mid X = 1) = 0$ and $P_1(Y = 1 \mid X = 1) = 1$, with both having $P(Y = 1 \mid X = 0) = 0$. Then $d_{TV}(P_0, P_1) = \eta$ but $d_{TV}(\Phi_1(P_0), \Phi_1(P_1)) = 1$. Let $\mathcal{P}_\eta = \operatorname{conv}\{P_0, P_1\}$. Then $\operatorname{diam}_{TV}(\mathcal{P}_\eta) = \eta$, $\alpha_{\min}(\mathcal{P}_\eta) = \eta$, and $W(\mathcal{P}_\eta, 1) = 1$. □

5.3 COMPUTATIONAL COMPLEXITY

Proposition 18 (Bilinear Program for W). *If $\mathcal{P}$ is specified by linear constraints on conditional probability table (CPT) blocks, then computing W is a bilinear program. If $\mathcal{P}$ is specified by constraints on the joint distribution, the problem involves a nonconvex sum-of-fractions optimization; see Appendices D.1 and D.2.*

Theorem 19 (NP-hardness of W; see Appendix D.3). *Deciding whether $W(\mathcal{P}, x) \geq \tau$ for a rational threshold τ is NP-hard, even for binary Y, trivial X, and $K \geq 2$ strata. The proof proceeds by a polynomial-time reduction from maximum clique using the Motzkin–Straus theorem.*

This intractability clarifies why exact robust identification widths are not merely a matter of evaluating a closed-form formula. It is consistent with the broader literature on credal network inference, where exact bounding tasks are NP-hard even for restricted graph classes [Mauá et al., 2013].

6 MINIMAX REGRET FOR CAUSAL DECISIONS

We now establish decision-theoretic consequences of credal uncertainty. The estimation and regret results use the same principle: two observationally plausible worlds can have separated interventional or decision-relevant quantities.

6.1 SETUP

Definition 20 (Causal Decision Problem). A causal decision problem consists of an action space $\mathcal{A}$ of possible interventions, an outcome space $\mathcal{Y}$, a utility function $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, and an unknown true distribution $P^* \in \mathcal{P}$. The expected utility of action a under P is $U_P(a) = \mathbb{E}_{P(Y|\mathrm{do}(a))}[u(a, Y)]$, and an optimal action is $a^*(P) \in \arg\max_{a \in \mathcal{A}} U_P(a)$.

Definition 21 (Regret). The regret of action a under distribution P is

$$R(a, P) = U_P(a^*(P)) - U_P(a). \tag{15}$$

For a randomized decision rule $\pi \in \Delta(\mathcal{A})$, define $U_P(\pi) = \mathbb{E}_{a \sim \pi}[U_P(a)]$ and $R(\pi, P) = U_P(a^*(P)) - U_P(\pi)$.

For meaningful regret lower bounds, we require a non-degeneracy condition ensuring that no single action hedges well across the least-favorable worlds.

Definition 22 (Strongly Action-Separating Credal Set). For $P \in \mathcal{P}$, let $a_P^\star \in \arg\max_{a \in \mathcal{A}} V(a; P)$ and $R(a, P) = V(a_P^\star; P) - V(a; P)$. A credal set $\mathcal{P}$ is (g, P_0, P_1)-strongly action-separating if there exist two worlds $P_0, P_1 \in \mathcal{P}$ such that, for every pure action $a \in \mathcal{A}$,

$$\max\{R(a, P_0), R(a, P_1)\} \geq g.$$

Here P_0, P_1 are distributions, not actions.

This condition is stronger than merely requiring different optimal actions under P_0 and P_1. A weaker condition would not suffice because a third action might achieve low regret under both worlds. The condition is therefore a least-favorable decision construction, not a claim that all credal decision problems are strongly separated.

6.2 PURE MINIMAX REGRET

Theorem 23 (Minimax Regret Lower Bound, Pure Actions). *Let $\mathcal{P}$ be a (g, P_0, P_1)-strongly action-separating credal set. Then*

$$\min_{a \in \mathcal{A}} \max_{P \in \mathcal{P}} R(a, P) \geq g. \tag{16}$$

Proof. By strong action-separation, for every $a \in \mathcal{A}$, $\max\{R(a, P_0), R(a, P_1)\} \geq g$. Since $P_0, P_1 \in \mathcal{P}$, $\max_{P \in \mathcal{P}} R(a, P) \geq g$ for all a, and minimizing over a preserves the bound. □

6.3 RANDOMIZED MINIMAX REGRET

When randomized decisions are allowed, the minimax regret can be reduced in the symmetric least-favorable construction. The factor two is a hedging effect: if action a_0 is better by g in world P_0 and action a_1 is better by g in world P_1, any deterministic rule commits to one action and suffers regret g in the opposite world. A randomized rule that chooses each action with probability $1/2$ suffers regret $g/2$ in both worlds. This does not mean arbitrary randomization always halves regret; the statement relies on the symmetric action-separation assumptions below.

Theorem 24 (Minimax Regret with Randomization)**.** *Let $\mathcal{P}$ contain P_0, P_1 such that: (i) $a_0 = a^*(P_0) \neq a^*(P_1) = a_1$; (ii) $U_{P_0}(a_0) - U_{P_0}(a_1) = g$ and $U_{P_1}(a_1) - U_{P_1}(a_0) = g$; and (iii) for all $a \notin \{a_0, a_1\}$, $R(a, P_0) \geq g$ and $R(a, P_1) \geq g$. Then*

$$\min_{\pi \in \Delta(\mathcal{A})} \max_{P \in \mathcal{P}} R(\pi, P) = \frac{g}{2}, \tag{17}$$

achieved by $\pi^* = \frac{1}{2}\delta_{a_0} + \frac{1}{2}\delta_{a_1}$.

Proof sketch. For any π, let $p_0 = \pi(a_0)$ and $p_1 = \pi(a_1)$. Under P_0, not playing a_0 incurs regret at least g, so $R(\pi, P_0) \geq (1 - p_0)g$. Under P_1, $R(\pi, P_1) \geq (1 - p_1)g$. Hence the worst-case regret is at least $\max\{(1 - p_0)g, (1 - p_1)g\}$, which is minimized at $p_0 = p_1 = 1/2$. The policy π^* attains $g/2$ in both worlds. The full proof is in Appendix F.6. □

Corollary 25 (Pure vs. Randomized Gap)**.** *Under the conditions of Theorems 23 and 24, the pure minimax regret is g and the randomized minimax regret is $g/2$. If no strongly action-separating pair exists, for example when all distributions in $\mathcal{P}$ share the same optimal action, minimax regret can be small even for large credal diameter.*

7 PRACTICAL PROCEDURES

The optimization problems introduced in Sections 5 and 6 directly yield practical procedures for (i) constructing credal-robust causal effect intervals and (ii) selecting actions under distributional ambiguity.

We consider two estimation settings. When the credal set is specified a priori, causal effects are obtained by optimizing the causal functional over the admissible distribution set. When the credal set is constructed from data, finite-sample guarantees follow from total-variation concentration. For decision making, we implement a conservative minimax-regret rule based on lower and upper bounds on interventional utilities.

For completeness, detailed pseudocode for all procedures is provided in Appendix B.

8 EXPERIMENTS

All experiments are diagnostic checks of the main scaling laws using synthetic discrete SCMs. They are not intended as broad empirical benchmarks or as evidence that the hard pairs arise automatically in realistic causal workflows. Instead, we instantiate the least-favorable constructions from the proofs, where the bounds are expected to be tight, and verify the predicted dependence on positivity, accuracy, strata count, robust width, and regret. We define the empirical sample complexity n^* as the smallest n such that the worst-case error probability over the evaluated distributions is at most δ. Monte Carlo quantities are averaged over repeated trials; deterministic computations, such as exact minimax regret, are reported without uncertainty bands.

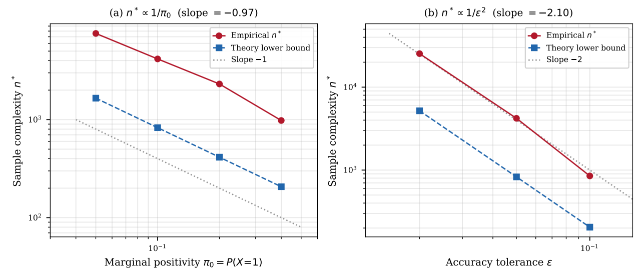

Figure 1: Sample complexity validates Proposition 2. (a) Log-log plot of empirical n^* vs. marginal positivity π_0; fitted slope -0.97 matches theory (-1). (b) Log-log plot of n^* vs. accuracy ε; fitted slope -2.10 matches theory (-2). Dashed lines show theoretical lower bounds; dotted lines show reference slopes.

8.1 EXPERIMENT 1: SAMPLE COMPLEXITY SCALING (PROPOSITION 2)

We validate the sample complexity scaling $n^* = \Omega(1/(\pi_0\varepsilon^2))$ using the symmetric Bernoulli hard pair from Proposition 2. The design uses binary (X, Y) with no confounding: $P(X = x) = \pi_0$, $P_0(Y = 1 \mid X = x) = 1/2 - \varepsilon$, $P_1(Y = 1 \mid X = x) = 1/2 + \varepsilon$, and P_0, P_1 identical elsewhere. We sweep $\pi_0 \in \{0.05, 0.1, 0.2, 0.4\}$ at fixed $\varepsilon = 0.05$ and $\delta = 0.05$, and sweep $\varepsilon \in \{0.02, 0.05, 0.1\}$ at fixed $\pi_0 = 0.1$ and $\delta = 0.05$, using the plug-in estimator $\hat{p} = \hat{P}(Y = 1 \mid X = x)$ over 500 trials per configuration. Figure 1 shows log-log plots of empirical sample complexity against marginal positivity and accuracy. The fitted slopes (-0.97 and -2.10) closely match the theoretical predictions (-1 and -2), confirming the propensity-driven scaling. The empirical n^* exceeds the lower bound by a constant factor, as expected for a Le Cam lower bound.

8.2 EXPERIMENT 2: ATE VS. UNIFORM ESTIMATION (COROLLARY 6)

A natural question is whether sample complexity scales with the number of strata K. Figure 2 resolves this in the least-favorable adjustment construction: for ATE estimation, the fitted slope is approximately 0.11, while for uniform conditional estimation, the fitted slope is approximately 1.26. The design uses a fork graph ($Z \to X$, $Z \to Y$, $X \to Y$) with $K \in \{2, 4, 8, 16, 32\}$ strata, equal-mass constant-propensity $P(Z = z) = 1/K$ and $P(X = x \mid Z = z) = 0.2$, and $\varepsilon = \delta = 0.05$ over 200 trials per K; see Table 1. This validates the distinction in Corollary 6: the ATE averages over strata, so rare strata receive small weight, whereas uniform estimation requires adequate samples in every stratum.

Table 1: Sample complexity for ATE vs. uniform conditional estimation as a function of strata count K. ATE remains approximately constant, while uniform estimation scales nearly linearly with K.

K	n^*_{ATE}	n^*_{Uniform}	Ratio
2	1,837	4,867	2.6
4	1,460	9,324	6.4
8	1,536	26,202	17.1
16	1,779	57,710	32.4
32	2,443	151,590	62.1

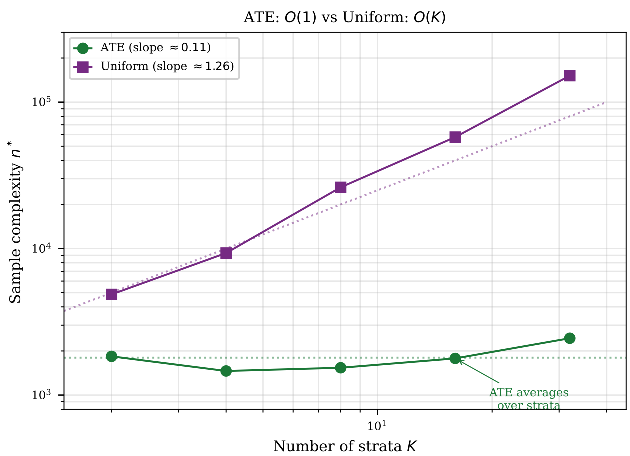

Figure 2: ATE sample complexity does not scale with K in the constant-propensity adjustment construction. ATE averages over strata, so rare strata contribute less; uniform conditional estimation requires adequate samples in every stratum.

8.3 EXPERIMENT 3: ROBUST IDENTIFICATION WIDTH (THEOREM 15)

We validate the Lipschitz bound $W(\mathcal{P}, x) \leq \min\{1, (1 + 1/\alpha_{\min})\Delta\}$ in two positivity regimes. The design uses a fork graph with adjustment, $K = 4$ strata, and a TV-ball credal set of radius $r = \Delta/2 = 0.05$ around a base distribution P_0 (so $\Delta = 0.10$). Positivity is controlled by choosing $\min_z P_0(X = x, Z = z) = \alpha_{\min} + r$. We approximate W by sampling 500 pairs $(P, Q) \in \mathcal{P}$ with $d_{TV}(P, Q) \approx \Delta$ and computing $d_{TV}(\Phi_x(P), \Phi_x(Q))$. Under strong positivity ($\alpha_{\min} = 0.15$), the theoretical bound is 0.77 and the empirical width is 0.07. Under weak positivity ($\alpha_{\min} = 0.02$), the theoretical bound saturates at 1.0 and the empirical width is 0.22. The gap is expected: Theorem 15 is a worst-case upper bound, and random search need not find adversarial pairs that achieve it. In implementations with random search, this step can be repeated over multiple seeds to report variability of the approximation to W.

8.4 EXPERIMENT 4: MINIMAX REGRET GAP (THEOREMS 23/24)

We verify the pure vs. randomized regret gap under strong action separation using exact computation rather than Monte Carlo. We construct two actions $\{a_0, a_1\}$ with binary outcome and utility $u(a, y) = \mathbb{1}[y = 1]$, with $U_{P_0}(a_0) = 1$, $U_{P_0}(a_1) = 1 - g$, $U_{P_1}(a_1) = 1$, $U_{P_1}(a_0) = 1 - g$, and vary $g \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For all g, the pure minimax regret equals g, while the randomized minimax regret at $\pi^* = (1/2, 1/2)$ equals $g/2$, yielding a constant ratio of 2. This confirms that randomization halves worst-case regret in the symmetric least-favorable construction; it is not a universal claim about arbitrary credal decision problems.

8.5 SUMMARY

Table 2 summarizes the diagnostic validation. The experiments test the theory at its tightest points rather than on arbitrary real-data instances. All predicted scalings are observed: sample complexity follows $1/(\pi_0\varepsilon^2)$, ATE has no K-dependence while uniform estimation scales with K, the Lipschitz bound holds with the predicted positivity amplification, and randomized policies achieve the factor-two regret improvement under strong action-separation.

9 DISCUSSION

Tightness and upper bounds. Empirical sample complexity exceeds the lower bound by a constant factor, as expected for Le Cam–type lower bounds. The matching upper bounds in Appendix C show that the rates are tight in the main special cases: the plug-in estimator achieves $O(1/(\pi_{\min}\varepsilon^2)\log(1/\delta))$ without confounding, and IPW achieves $O(\kappa(\mathcal{P})\varepsilon^{-2}\log(1/\delta))$ with known propensity.

What is standard, and what is causal. The information-theoretic tools used in the lower bounds are standard. The causal contribution is the identification of how these tools behave after the observational distribution is mapped to an interventional quantity. The same observational TV or KL diameter can lead to different causal difficulty depending on treatment propensity and adjustment weights.

Computation. Exact computation of $W(\mathcal{P}, x)$ is NP-hard (Theorem 19). Practical approximations include convex relaxations, alternating maximization, and mixed-integer formulations for small instances. This computational result explains why robust identification width is nontrivial to compute even in discrete settings.

Scope and extensions. The main concrete results are for finite discrete variables, with binary outcomes in the sharp hard-pair constructions and simple adjustment structures.

Table 2: Experimental validation of the theoretical predictions. Slopes are fitted by log-log regression. Sampling-based quantities are computed from repeated trials; exact regret calculations are deterministic.

Result	Prediction	Empirical	Status
Prop. 2	$n^* \propto 1/\pi_0$	slope $= -0.97$	✓
Prop. 2	$n^* \propto 1/\varepsilon^2$	slope $= -2.10$	✓
Cor. 6	ATE: $O(1)$ in K	slope $= 0.11$	✓
Cor. 6	Uniform: $O(K)$	slope $= 1.26$	✓
Thm. 15	$W \leq (1 + 1/\alpha_{\min})\Delta$	$0.07 \leq 0.77$, $0.22 \leq 1.0$	✓
Thm. 23/24	Pure/Randomized $= 2$	ratio $= 2.00$	✓

For outcomes with more than two categories, the binary construction can be embedded into two categories. Continuous outcomes, continuous confounders, and general DAGs require additional assumptions such as density-ratio control, smoothness, or kernel-based uncertainty sets.

Credal set learning. We assume the credal set is given. This abstraction is appropriate for studying information-theoretic limits, but practical causal workflows must estimate or elicit $\mathcal{P}$. Credal sets may come from finite-sample confidence regions, multi-domain data, covariate-shift neighborhoods, partial mechanism constraints, missing-data assumptions, or sensitivity-analysis models. Developing statistically valid methods for learning such sets is an important direction for future work.

10 CONCLUSION

We have studied causal inference when the observational distribution is known only through a credal set. The central message is that causal uncertainty is not governed by credal diameter alone: identification formulas can amplify or attenuate observational uncertainty depending on positivity. In the discrete hard-pair constructions, sample complexity scales as $\Omega(1/(\pi_0\varepsilon^2)\log(1/\delta))$ or $\Omega(1/(\beta_0\varepsilon^2)\log(1/\delta))$, and ATE estimation avoids the additional K-scaling that appears in uniform conditional estimation. Robust identification width is controlled by a Lipschitz factor $(1 + 1/\alpha_{\min})$, while causal decision-making under strongly separated credal ambiguity admits a factor-two improvement from randomization. Together, these results provide information-theoretic and decision-theoretic foundations for studying how observational credal uncertainty propagates through causal maps.

References

Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to Imprecise Probabilities*. John Wiley & Sons, June 2014. ISBN 978-0-470-97381-3. Google-Books-ID: 9qXIEAAAQBAJ.

Damiano Azzolini and Fabrizio Riguzzi. Inference in probabilistic answer set programs with imprecise probabilities via optimization. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, July 2016. doi: 10.1073/pnas.1510507113.

Juraj Bodik and Valérie Chavez-Demoulin. Identifiability of causal graphs under non-additive conditionally parametric causal models. *Journal of Machine Learning Research*, 26(264):1–55, 2025.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York, NY, 1986. ISBN 978-1-4612-9369-9 978-1-4612-4946-7. doi: 10.1007/978-1-4612-4946-7.

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.

Spencer Compton, Kristjan Greenewald, Dmitriy A Katz, and Murat Kocaoglu. Entropic Causal Inference: Graph Identifiability. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4311–4343. PMLR, July 2022.

Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(00)00029-1.

Abhinandan Dalal and Eric J. Tchetgen Tchetgen. Partial identification of causal effects for endogenous continuous treatments, 2025.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, June 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2004.

Noam Finkelstein, Roy Adams, Suchi Saria, and Ilya Shpitser. Partial identifiability in discrete data with measurement error. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1798–1808. PMLR, 27–30 Jul 2021.

Justus Hibshman and Tim Weninger. Higher order imprecise probabilities and statistical testing, 2021.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459. doi: 10.2307/2282952.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952. ISSN 0162-1459. doi: 10.2307/2280784.

Peter J. Huber and Volker Strassen. Minimax Tests and the Neyman-Pearson Lemma for Capacities. *The Annals of Statistics*, 1(2):251–263, March 1973. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342363.

Nathan Huntley, Robert Hable, and Matthias C. M. Troffaes. Decision making. In *Introduction to Imprecise Probabilities*, pages 190–206. John Wiley & Sons, Ltd, 2014. ISBN 978-1-118-76311-7. doi: 10.1002/9781118763117.ch8.

Murat Kocaoglu, Sanjay Shakkottai, Alexandros G. Dimakis, Constantine Caramanis, and Sriram Vishwanath. Applications of common entropy for causal inference. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

L. LeCam. Convergence of Estimates Under Dimensionality Restrictions. *The Annals of Statistics*, 1 (1):38–53, January 1973. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1193342380.

Sanghack Lee and Elias Bareinboim. Causal effect identifiability under partial-observability. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5692–5701. PMLR, 13–18 Jul 2020.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Science & Business Media, May 2003. ISBN 978-0-387-00454-9. Google-Books-ID: ESws8CEGW6AC.

Denis D. Mauá, Cassio P. de Campos, Alessio Benavoli, and Alessandro Antonucci. On the complexity of strong and epistemic credal networks. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 391–400, Arlington, Virginia, USA, 2013. AUAI Press.

Mariela Morveli-Espinoza, Juan Carlos Nieves, and Cesar Augusto Tacla. An imprecise probability approach for abstract argumentation based on credal sets. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 15th European Conference, ECSQARU 2019, Belgrade, Serbia, September 18-20, 2019, Proceedings*, page 39–49, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-29764-0.

T. S. Motzkin and E. G. Straus. Maxima for Graphs and a New Proof of a Theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540, January 1965. ISSN 0008-414X, 1496-4279. doi: 10.4153/CJM-1965-053-6.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, March 2000. ISBN 978-0-521-77362-1. Google-Books-ID: wnGU_TsW3BQC.

Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, New York, NY, 2002. ISBN 978-1-4419-3191-7 978-1-4757-3692-2. doi: 10.1007/978-1-4757-3692-2.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1219–1226. AAAI Press, 2006. ISBN 978-1-57735-281-5.

Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2006.06.001.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Taylor & Francis, 1991. ISBN 978-0-412-28660-5.

Kun Zhang, Jiji Zhang, Biwei Huang, Bernhard Schölkopf, and Clark Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 825–834, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

Qingyuan Zhao, Dylan S. Small, and Bhaswar B. Bhattacharya. Sensitivity Analysis for Inverse Probability Weighting Estimators via the Percentile Bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, September 2019. ISSN 1369-7412. doi: 10.1111/rssb.12327.

Information-Theoretic Lower Bounds for Causal Inference under Credal Uncertainty (Supplementary Material)

Hung Mai[1,2] **Hai Nguyen**[3] **Khanh Nguyen**[*1,3] **Luong Doan**[1] **Nhung Duong**[1] **Phong Ho**[1] **Tuan Do**[1,3]

[1]N2TP Technology Solutions JSC, Vietnam
[2]National Economics University, Vietnam
[3]Phenikaa University, Vietnam

A BACKGROUND DEFINITIONS

A.1 STRUCTURAL CAUSAL MODELS

A structural causal model (SCM) is a tuple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}))$ with endogenous variables $\mathbf{V}$, exogenous variables $\mathbf{U}$, and structural equations $V_i = f_i(\mathrm{Pa}_i, U_i)$ defining a causal graph.

An intervention $\mathrm{do}(X = x)$ replaces the structural equation of X by the constant x, producing a modified model $\mathcal{M}_x$ with interventional distribution $P_{\mathcal{M}_x}$.

If a backdoor adjustment set $\mathbf{Z}$ exists,

$$P(Y \mid \mathrm{do}(X = x)) = \sum_{\mathbf{z}} P(Y \mid X = x, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z}). \tag{18}$$

A.2 CREDAL SETS

A credal set $\mathcal{P}$ over a finite sample space Ω is a nonempty closed convex set of probability distributions.

Lower and upper probabilities are

$$\underline{P}(A) = \inf_{P \in \mathcal{P}} P(A), \qquad \overline{P}(A) = \sup_{P \in \mathcal{P}} P(A).$$

A.3 INFORMATION THEORY

The KL divergence is

$$D_{KL}(P \| Q) = \sum_{\omega} P(\omega) \log \frac{P(\omega)}{Q(\omega)}.$$

We use the Bretagnolle–Huber inequality

$$d_{TV}(P^n, Q^n)^2 \leq 1 - \exp(-n D_{KL}(P \| Q)).$$

B ALGORITHMS AND PRACTICAL PROCEDURES

We present algorithms for estimating causal effects and making decisions under credal uncertainty. These procedures implement the optimization problems introduced in the main text and provide deterministic or finite-sample guarantees depending on how the credal set is specified.

*Correspondence: khanh.nguyenquoc@phenikaa-uni.edu.vn
*Correspondence: khanh.nguyenquoc@phenikaa-uni.edu.vn

B.1 CREDAL-ROBUST CAUSAL EFFECT ESTIMATION

We consider two settings depending on whether the credal set is fixed a priori or constructed from data.

Algorithm 1 Robust Interval under Fixed Credal Set

Require: Credal set $\mathcal{P}$ (fixed and known a priori), treatment value x, adjustment set Z
Ensure: Interval $[\underline{\theta}, \overline{\theta}]$ for $\mathbb{E}[Y \mid \mathrm{do}(X = x)]$
Compute $\underline{\theta} = \inf_{P \in \mathcal{P}} \theta(P)$
Compute $\overline{\theta} = \sup_{P \in \mathcal{P}} \theta(P)$
return $[\underline{\theta}, \overline{\theta}]$

If the true distribution P^* belongs to $\mathcal{P}$, then the resulting interval contains the true causal effect deterministically:

$$\theta(P^*) \in [\underline{\theta}, \overline{\theta}].$$

When $\mathcal{P}$ is specified by constraints on joint probabilities and adjustment is required, the optimization involves a sum-of-fractions objective (Appendix D.2). Practical approaches include bilinearization using auxiliary conditional variables or mixed-integer reformulations.

Algorithm 2 Statistical Credal Interval

Require: Dataset $D = \{(x_i, y_i, z_i)\}_{i=1}^n$, confidence $1 - \delta$, treatment value x, adjustment set Z, sample space size $m = |\Omega|$
Ensure: $(1 - \delta)$ confidence interval $[\underline{\theta}_n, \overline{\theta}_n]$
Compute empirical distribution $\hat{P}$ from D
Set $r_n(\delta) = \sqrt{\log((2^m - 2)/\delta)/(2n)}$
Construct $\mathcal{P}_n(\delta) = \{P : d_{TV}(P, \hat{P}) \leq r_n(\delta)\}$
Compute $\underline{\theta}_n = \inf_{P \in \mathcal{P}_n(\delta)} \theta(P)$
Compute $\overline{\theta}_n = \sup_{P \in \mathcal{P}_n(\delta)} \theta(P)$
return $[\underline{\theta}_n, \overline{\theta}_n]$

By the finite-alphabet total-variation concentration bound of Weissman et al.,

$$\Pr\Big(d_{TV}(\hat{P}_n, P^*) \leq r_n(\delta)\Big) \geq 1 - \delta.$$

Conditioned on this event, $P^* \in \mathcal{P}_n(\delta)$ and therefore $\theta(P^*) \in [\underline{\theta}_n, \overline{\theta}_n]$. Algorithm 2 assumes a finite sample space; continuous settings require Wasserstein balls or kernel-based uncertainty sets.

B.2 CONSERVATIVE MINIMAX REGRET DECISION RULE

Algorithm 3 Conservative Minimax Regret Upper Bound Rule

Require: Action space $\mathcal{A}$, credal set $\mathcal{P}$, utility u
Ensure: Action a^*_{CMR}
for each $a \in \mathcal{A}$ **do**
  Compute $U_L(a) = \inf_{P \in \mathcal{P}} \mathbb{E}_{P(Y|\mathrm{do}(a))}[u(a, Y)]$
  Compute $U_U(a) = \sup_{P \in \mathcal{P}} \mathbb{E}_{P(Y|\mathrm{do}(a))}[u(a, Y)]$
  Compute $\overline{R}(a) = \max_{a' \in \mathcal{A}} U_U(a') - U_L(a)$
end for
return $a^*_{\mathrm{CMR}} = \arg\min_{a \in \mathcal{A}} \overline{R}(a)$

This rule is conservative because the supremum and infimum may be attained at different distributions. For all actions a,

$$\max_{P \in \mathcal{P}} R(a, P) \leq \overline{R}(a).$$

When the credal set is polytopic over interventional distributions, exact minimax regret can be computed by solving $|\mathcal{A}|$ linear programs (Appendix F.7).

C MATCHING UPPER BOUNDS FOR SAMPLE COMPLEXITY

The main text establishes information-theoretic lower bounds of order $n = \Omega(1/(\text{positivity} \cdot \varepsilon^2) \log(1/\delta))$ in the binary hard-pair setting (for $\delta < 1/2$). Here we prove matching achievability upper bounds up to constants.

C.1 NO-CONFOUNDING CASE

Setting. Assume (X, Y) are observed i.i.d. from some unknown $P \in \mathcal{P}$. The causal estimand is identifiable without adjustment: $\theta(P) = \mathbb{E}_P[Y \mid \mathrm{do}(X = x)] = \mathbb{E}_P[Y \mid X = x]$. Assume $Y \in [0, 1]$ and a uniform positivity lower bound $\pi_{\min} := \inf_{P \in \mathcal{P}} P(X = x) > 0$. Define the plug-in conditional mean estimator:

$$\widehat{\theta} := \begin{cases} \frac{1}{N_x} \sum_{i=1}^n Y_i \mathbf{1}\{X_i = x\}, & N_x > 0 \\ 0, & N_x = 0 \end{cases} \quad \text{where } N_x := \sum_{i=1}^n \mathbf{1}\{X_i = x\} \tag{19}$$

Theorem 26 (Achievability; no confounding). *For any $\varepsilon \in (0, 1]$ and $\delta \in (0, 1)$,*

$$\sup_{P \in \mathcal{P}} P^n\Big(|\widehat{\theta} - \theta(P)| > \varepsilon\Big) \le \delta \tag{20}$$

whenever

$$n \ge \max\left\{\frac{8}{\pi_{\min}} \log \frac{2}{\delta},\ \frac{1}{\pi_{\min}\varepsilon^2} \log \frac{4}{\delta}\right\} \tag{21}$$

In particular, $n = O\Big(\frac{1}{\pi_{\min}\varepsilon^2} \log \frac{1}{\delta}\Big)$.

Proof. Fix any $P \in \mathcal{P}$. Let $\pi := P(X = x) \ge \pi_{\min}$.

Step 1: Ensure enough treated samples. $N_x \sim \mathrm{Bin}(n, \pi)$. By the multiplicative Chernoff bound (with deviation $1/2$),

$$P\Big(N_x \le \frac{n\pi}{2}\Big) \le \exp\Big(-\frac{n\pi}{8}\Big) \tag{22}$$

Since $\pi \ge \pi_{\min}$, we have $\{N_x \le n\pi_{\min}/2\} \subseteq \{N_x \le n\pi/2\}$. Therefore

$$P\Big(N_x \le \frac{n\pi_{\min}}{2}\Big) \le \exp\Big(-\frac{n\pi_{\min}}{8}\Big) \tag{23}$$

Step 2: Conditional concentration given N_x. Conditional on $N_x = m > 0$, the m observed outcomes with $X_i = x$ are i.i.d. from $P(Y \mid X = x)$, bounded in $[0, 1]$, with mean $\theta(P)$, and $\widehat{\theta}$ is their sample mean. By Hoeffding's inequality [Hoeffding, 1963]:

$$P\Big(|\widehat{\theta} - \theta(P)| > \varepsilon \;\Big|\; N_x = m\Big) \le 2\exp(-2m\varepsilon^2) \tag{24}$$

On the event $N_x \ge n\pi_{\min}/2$, we have $m \ge n\pi_{\min}/2$, hence

$$2\exp(-2m\varepsilon^2) \le 2\exp(-n\pi_{\min}\varepsilon^2) \tag{25}$$

Step 3: Combine by the law of total probability.

$$P\Big(|\widehat{\theta} - \theta(P)| > \varepsilon\Big) \le P\Big(N_x < \frac{n\pi_{\min}}{2}\Big) + P\Big(|\widehat{\theta} - \theta(P)| > \varepsilon,\ N_x \ge \frac{n\pi_{\min}}{2}\Big) \tag{26}$$

$$\le \exp\Big(-\frac{n\pi_{\min}}{8}\Big) + 2\exp(-n\pi_{\min}\varepsilon^2) \tag{27}$$

Choose n so that each term is $\le \delta/2$. Since the bound depends only on $\pi_{\min}$, taking $\sup_{P \in \mathcal{P}}$ preserves it. □

In the symmetric hard-pair construction (Proposition 2), the KL scales as $\pi_0\varepsilon^2$, giving $n = \Omega(1/(\pi_0\varepsilon^2)\log(1/\delta))$. Theorem 26 shows a matching upper bound when $\pi_{\min}$ plays the role of π_0.

C.2 ADJUSTMENT CASE WITH IPW

Setting. Assume back-door adjustment is valid with covariates Z, and there exists a known function $e : \mathcal{Z} \to (0, 1]$ such that for every $P \in \mathcal{P}$, $P(X = x \mid Z = z) = e(z)$ for all z. Assume $Y \in [0, 1]$, and define $\beta_{\min} := \inf_z e(z) > 0$ and $\kappa(\mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[1/e(Z)]$. Use the IPW estimator:

$$\widehat{\theta}_{IPW} := \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{X_i = x\} Y_i}{e(Z_i)} \tag{28}$$

Theorem 27 (Achievability; adjustment via IPW)**.** *Under the known-propensity assumption, for any $\varepsilon > 0$ and $\delta \in (0, 1)$,*

$$\sup_{P \in \mathcal{P}} P^n \left(|\widehat{\theta}_{IPW} - \theta(P)| > \varepsilon \right) \le \delta \tag{29}$$

whenever

$$n \ge \left(\frac{2\kappa(\mathcal{P})}{\varepsilon^2} + \frac{2}{3} \frac{1}{\beta_{\min}} \frac{1}{\varepsilon} \right) \log \frac{2}{\delta} \tag{30}$$

For the usual accuracy regime $\varepsilon \ll 1$, the $1/\varepsilon^2$ term dominates, giving $n = O(\kappa(\mathcal{P})/\varepsilon^2 \cdot \log(1/\delta))$.

Proof. Fix $P \in \mathcal{P}$. Define $W_i := \mathbf{1}\{X_i = x\} Y_i / e(Z_i)$. Then $\widehat{\theta}_{IPW} = \frac{1}{n} \sum_i W_i$.

Step 1: Unbiasedness. Conditioning on $Z = z$, since $\mathbf{1}\{X = x\}$ is Bernoulli with mean $e(z)$:

$$\mathbb{E}_P[W_i \mid Z = z] = \frac{1}{e(z)} \cdot e(z) \cdot \mathbb{E}_P[Y \mid X = x, Z = z] = \mathbb{E}_P[Y \mid X = x, Z = z] \tag{31}$$

Taking expectation over Z yields $\mathbb{E}_P[W_i] = \theta(P)$.

Step 2: Almost-sure bound. Since $Y \in [0, 1]$ and $e(Z) \ge \beta_{\min}$, $0 \le W_i \le 1/\beta_{\min} =: B$.

Step 3: Variance bound reveals κ. Because $0 \le Y \le 1$,

$$\mathbb{E}_P[W_i^2 \mid Z] \le \frac{e(Z)}{e(Z)^2} = \frac{1}{e(Z)} \tag{32}$$

So $\mathrm{Var}_P(W_i) \le \mathbb{E}_P[W_i^2] \le \mathbb{E}_P[1/e(Z)] \le \kappa(\mathcal{P}) =: \sigma^2$.

Step 4: Apply Bernstein's inequality. Let $\widetilde{X}_i := W_i - \mathbb{E}[W_i]$. Then $|\widetilde{X}_i| \le B$ and $\sum_i \mathbb{E}[\widetilde{X}_i^2] \le n\sigma^2$. Bernstein's inequality gives:

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_i \right| > \varepsilon \right) \le 2 \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3} B\varepsilon} \right) \tag{33}$$

Setting the RHS $\le \delta$ and solving for n gives the theorem. $\square$

In the equal-mass constant-propensity example from Remark 3.3 of the main text, $e(Z) \equiv \beta$, hence $\kappa = 1/\beta$, independent of K.

Remark (Unknown propensity). If the propensity $e_P(\cdot)$ varies across $P \in \mathcal{P}$ and is unknown, the IPW estimator above is not implementable. Achievability in the minimax sense with unknown propensity would require analyzing doubly-robust or cross-fitted estimators; we defer this to future work.

C.3 STANDARD INEQUALITIES

Two key ingredients used in the lower-bound machinery (Section 4) are standard:

The inequality $d_{TV}(P, Q) \le \sqrt{1 - \exp(-D_{KL}(P \| Q))}$ is the Bretagnolle-Huber inequality, commonly used for Le Cam testing lower bounds.

The finite-alphabet concentration used in Algorithm 2, $\Pr(\|\widehat{P}_n - P\|_1 > \varepsilon) \le (2^m - 2) \exp(-n\varepsilon^2/2)$, is (up to constants) the Weissman-Ordentlich-Seroussi-Verdu-Weinberger bound for the ℓ_1 deviation of the empirical distribution.

D COMPUTATIONAL COMPLEXITY OF $W(\mathcal{P}, x)$

D.1 BILINEAR PROGRAM UNDER CPT-BLOCK CONSTRAINTS

Assume Y and Z are finite and the causal effect is identifiable via back-door adjustment. Suppose $\mathcal{P}$ is specified by linear constraints on the CPT blocks $p_z := P(Z = z)$ and $m_{yz} := P(Y = y \mid X = x, Z = z)$, together with simplex constraints. We identify each $P \in \mathcal{P}$ with its admissible CPT blocks (p, m), so $\mathcal{P}$ denotes the feasible polytope in block space.

Proposition 28 (Bilinear program for W). *Under the CPT-block specification,*

$$W(\mathcal{P}, x) = \max_{(p,m)\in\mathcal{P},\,(q,n)\in\mathcal{P}} \frac{1}{2} \sum_y \left| \sum_z p_z m_{yz} - \sum_z q_z n_{yz} \right| \tag{34}$$

This is a nonconvex (piecewise bilinear) program because the expressions $p_z m_{yz}$ and $q_z n_{yz}$ multiply decision variables, and the objective involves absolute values of these bilinear expressions. Since $\mathcal{P}$ is compact and the objective is continuous, the supremum is attained. The absolute values can be encoded exactly using binary variables: let $a_y := \sum_z p_z m_{yz} - \sum_z q_z n_{yz}$, with $a_y \in [-1, 1]$. The standard big-M encoding of $t_y = |a_y|$ requires $M = 2$ and binary variables $b_y \in \{0, 1\}$:

$$t_y \geq a_y, \quad t_y \geq -a_y, \quad t_y \leq a_y + 2(1 - b_y), \quad t_y \leq -a_y + 2b_y \tag{35}$$

Maximizing $\frac{1}{2}\sum_y t_y$ yields a mixed-integer bilinear program.

Proof. Start from the definition $W(\mathcal{P}, x) = \sup_{P,Q\in\mathcal{P}} d_{TV}(\Phi_x(P), \Phi_x(Q))$. Using the adjustment formula, $\Phi_x(P)(y) - \Phi_x(Q)(y) = \sum_z P(y \mid x, z)P(z) - \sum_z Q(y \mid x, z)Q(z)$. Renaming variables yields the stated expression. □

Because the objective is nonconvex in (p, m), the optimum need not occur at extreme points of the CPT-block polytope. For small instances, exact solutions can be obtained via the mixed-integer reformulation above or global nonconvex solvers; these are exponential in the worst case.

D.2 SUM-OF-FRACTIONS UNDER JOINT DISTRIBUTION CONSTRAINTS

When $\mathcal{P}$ is specified by linear constraints on the joint distribution $r_{y,x',z} = P(Y = y, X = x', Z = z)$, the interventional mapping involves ratios $P(y \mid x, z) = r_{y,x,z}/s_z$ where $s_z = \sum_{y'} r_{y',x,z}$. We assume uniform conditional positivity: there exists $\beta_{\min} > 0$ such that $P(X = x, Z = z) \geq \beta_{\min} P(Z = z)$ for all $P \in \mathcal{P}$ and all z. In joint-mass variables, this is the linear inequality $s_z \geq \beta_{\min} p_z$ where $p_z = P(Z = z)$.

Plugging into $\Phi_x(P)(y) = \sum_z P(y \mid x, z)\, P(z)$ yields

$$\Phi_x(P)(y) = \sum_{z:p_z>0} \frac{r_{y,x,z}}{s_z}\, p_z \tag{36}$$

a sum-of-fractions in the joint-mass variables. A standard bilinearization introduces conditional variables $m_{yz} := P(y \mid x, z) = r_{y,x,z}/s_z$ and constraints $r_{y,x,z} = m_{yz} \cdot s_z$, $s_z = \sum_{y'} r_{y',x,z}$, $\sum_{y'} m_{y'z} = 1$, $m_{yz} \geq 0$, and $s_z \geq \beta_{\min} p_z$, together with the joint-simplex constraints $r_{y,x',z} \geq 0$, $\sum_{y,x',z} r_{y,x',z} = 1$, and $p_z = \sum_{y',x'} r_{y',x',z}$. This recovers the CPT-block bilinear form of Proposition 28. The bilinear constraints $r_{y,x,z} = m_{yz} \cdot s_z$ are themselves nonconvex.

D.3 NP-HARDNESS OF COMPUTING W

Theorem 29 (NP-hardness of W). *Given a rational threshold $\tau \in \mathbb{Q}$, deciding whether $W(\mathcal{P}, x) \geq \tau$ is NP-hard, even in a very restricted discrete setting (binary Y, trivial X, $K \geq 2$ strata, and a credal set given by linear constraints on CPT blocks).*

Proof. We reduce from the maximum clique problem, which is NP-hard. By the Motzkin-Straus theorem [Motzkin and Straus, 1965], the clique number $\omega(G)$ of a graph G satisfies $\omega(G) = 1/(1 - \max_{p\in\Delta_K} p^\top Ap)$, where A is the adjacency matrix with zeros on the diagonal and Δ_K is the K-simplex.

Step 1: Construct a causal/credal instance. Given an undirected graph G with $K \geq 2$ vertices and adjacency matrix $A \in \{0,1\}^{K \times K}$ with zeros on the diagonal, define variables $Z \in \{1, \dots, K\}$, $X \in \{x\}$ (degenerate), and $Y \in \{0,1\}$. Define a credal set $\mathcal{P}$ over CPT blocks (p, m) with $p_i := P(Z = i)$ and $m_i := P(Y = 1 \mid X = x, Z = i)$, where p lies in the simplex and $m_i = \frac{1}{K-1} \sum_{j=1}^{K} A_{ij} p_j$. Since $A_{ij} \geq 0$ and row sums satisfy $\sum_j A_{ij} \leq K - 1$, we have $0 \leq m_i \leq 1$, so these constraints define a valid credal set.

Step 2: Compute the causal estimand. Since X is fixed to x, the adjustment formula gives $\theta(P) = \sum_{i=1}^{K} m_i p_i = \frac{1}{K-1} p^\top A p$.

Step 3: Relate W to the quadratic program. Because Y is binary, $d_{TV}(\Phi_x(P), \Phi_x(Q)) = |\theta(P) - \theta(Q)|$, so $W(\mathcal{P}, x) = \max_P \theta(P) - \min_P \theta(P)$. Since choosing $p = e_i$ (a vertex) gives $\theta = 0$ (because $A_{ii} = 0$), $\min_P \theta(P) = 0$, hence $W(\mathcal{P}, x) = \frac{1}{K-1} \max_{p \in \Delta_K} p^\top A p$.

The set of joint distributions induced by admissible CPT blocks need not be convex. As is standard in credal networks, define the credal set over joint distributions to be the convex hull (strong extension). Since $\theta(P) = P(Y = 1)$ is linear in the joint, the maximum and minimum over the convex hull equal those over the original set, so W is unchanged.

Step 4: Conclude NP-hardness. By the Motzkin-Straus theorem, $\max_{p \in \Delta_K} p^\top A p = 1 - 1/\omega(G)$. Therefore $\omega(G) \geq k$ if and only if $W(\mathcal{P}, x) \geq \frac{1}{K-1}(1 - 1/k)$. Given any clique instance (G, k), we construct the above credal set in polynomial time and check whether $W(\mathcal{P}, x) \geq \tau$ with $\tau = \frac{1}{K-1}(1 - 1/k)$. Since the clique problem is NP-complete, deciding $W(\mathcal{P}, x) \geq \tau$ is NP-hard. $\square$

E ADDITIONAL COROLLARIES

Corollary 30 (Conditional Positivity with Stratum Mass). *If we have conditional positivity $\beta_{\min} = \inf_P \min_z P(X = x \mid Z = z) > 0$ and stratum mass $\zeta_{\min} = \inf_P \min_z P(Z = z) > 0$, then $\alpha_{\min} \geq \zeta_{\min} \cdot \beta_{\min}$ and:*

$$W(\mathcal{P}, x) \leq \min\left\{1, \left(1 + \frac{1}{\zeta_{\min} \cdot \beta_{\min}}\right) \cdot \operatorname{diam}_{TV}(\mathcal{P})\right\} \tag{37}$$

F FULL PROOFS

F.1 PROOF OF LEMMA 1

By Le Cam's lemma, for any estimator $\hat{\theta}_n$:

$$\max_{i \in \{0,1\}} P_i^n\left(|\hat{\theta}_n - \theta(P_i)| > \varepsilon\right) \geq \frac{1 - d_{TV}(P_0^n, P_1^n)}{2} \tag{38}$$

For (ε, δ)-correctness, we need $d_{TV}(P_0^n, P_1^n) \geq 1 - 2\delta$. Since $\delta < 1/2$, the RHS is positive and squaring preserves the inequality.

Using the Bretagnolle-Huber inequality:

$$d_{TV}(P_0^n, P_1^n)^2 \leq 1 - \exp(-n \cdot D_{KL}(P_0 \| P_1)) \tag{39}$$

Requiring $d_{TV}(P_0^n, P_1^n) \geq 1 - 2\delta > 0$, we have $d_{TV}^2 \geq (1 - 2\delta)^2 = 1 - 4\delta(1 - \delta)$. Thus:

$$1 - \exp(-n \cdot D_{KL}(P_0 \| P_1)) \geq 1 - 4\delta(1 - \delta) \tag{40}$$

Solving:

$$n \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{D_{KL}(P_0 \| P_1)} \qquad \square \tag{41}$$

F.2 PROOF OF PROPOSITION 2

Step 1: KL divergence computation. Since P_0 and P_1 are identical except on $P(Y \mid X = x)$ (by assumption (iii)), the KL divergence localizes to the $X = x$ slice:

$$D_{KL}(P_0 \| P_1) = \pi_0 \cdot D_{KL}(\mathrm{Bern}(p_0) \| \mathrm{Bern}(p_1)) \tag{42}$$

Step 2: Symmetric case. For the symmetric pair $p_0 = \frac{1}{2} - \varepsilon$, $p_1 = \frac{1}{2} + \varepsilon$:

$$D_{KL}\Big(\mathrm{Bern}\big(\tfrac{1}{2} - \varepsilon\big) \,\Big\|\, \mathrm{Bern}\big(\tfrac{1}{2} + \varepsilon\big)\Big) = \big(\tfrac{1}{2} - \varepsilon\big) \log \frac{\frac{1}{2} - \varepsilon}{\frac{1}{2} + \varepsilon} + \big(\tfrac{1}{2} + \varepsilon\big) \log \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} = 2\varepsilon \log \frac{1 + 2\varepsilon}{1 - 2\varepsilon} \tag{43}$$

Step 3: Apply Lemma 1. By Lemma 1:

$$n \geq \frac{\log\Big(\frac{1}{4\delta(1-\delta)}\Big)}{\pi_0 \cdot 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}} \qquad \square \tag{44}$$

F.3 PROOF OF COROLLARY 6

Step 1: Verify ATE separation. Under the uniform shift construction:

$$\theta(P_0) = \sum_z P_0(Y{=}1 \mid x, z)\, P_0(z) = \big(\tfrac{1}{2} - \varepsilon\big) \sum_z P_0(z) = \tfrac{1}{2} - \varepsilon \tag{45}$$

$$\theta(P_1) = \sum_z P_1(Y{=}1 \mid x, z)\, P_1(z) = \big(\tfrac{1}{2} + \varepsilon\big) \sum_z P_1(z) = \tfrac{1}{2} + \varepsilon \tag{46}$$

So $|\theta(P_1) - \theta(P_0)| = 2\varepsilon$ as required.

Step 2: Compute KL divergence. Since P_0 and P_1 differ only in $P(Y \mid X{=}x, \mathbf{Z})$:

$$D_{KL}(P_0 \| P_1) = \sum_z P_0(X{=}x, \mathbf{Z}{=}z) \cdot D_{KL}\Big(\mathrm{Bern}\big(\tfrac{1}{2} - \varepsilon\big) \,\Big\|\, \mathrm{Bern}\big(\tfrac{1}{2} + \varepsilon\big)\Big) = P_0(X{=}x) \cdot 2\varepsilon \log \frac{1 + 2\varepsilon}{1 - 2\varepsilon} \tag{47}$$

Step 3: Apply Lemma 1.

$$n \geq \frac{\log\Big(\frac{1}{4\delta(1-\delta)}\Big)}{P_0(X{=}x) \cdot 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}} \tag{48}$$

If $P_0(X{=}x) = \beta_0$, then $n = \Omega(1/(\beta_0 \varepsilon^2) \log(1/\delta))$. $\square$

F.4 PROOF OF LEMMA 14

Assume without loss of generality that $P(B) \leq Q(B)$. Define $r := P(B)/Q(B) \in (0, 1]$.

For any atom $\omega \in B$:

$$P(\omega \mid B) - Q(\omega \mid B) = \frac{P(\omega)}{P(B)} - \frac{Q(\omega)}{Q(B)} = \frac{1}{P(B)} \left(P(\omega) - r \cdot Q(\omega)\right) \tag{49}$$

Therefore:

$$2\, d_{TV}(P(\cdot \mid B), Q(\cdot \mid B)) = \sum_{\omega \in B} |P(\omega \mid B) - Q(\omega \mid B)| = \frac{1}{P(B)} \sum_{\omega \in B} |P(\omega) - r \cdot Q(\omega)| \tag{50}$$

Since $\sum_{\omega \in B}(P(\omega) - r \cdot Q(\omega)) = P(B) - r \cdot Q(B) = 0$, the ℓ_1 norm equals twice the negative part:

$$\sum_{\omega \in B} |P(\omega) - r \cdot Q(\omega)| = 2 \sum_{\substack{\omega \in B: \\ P(\omega) < r \cdot Q(\omega)}} (r \cdot Q(\omega) - P(\omega)) \tag{51}$$

Since $r \leq 1$, we have $P(\omega) < r \cdot Q(\omega) \Rightarrow P(\omega) < Q(\omega)$. Therefore:

$$\sum_{\substack{\omega \in B: \\ P(\omega) < r \cdot Q(\omega)}} (r \cdot Q(\omega) - P(\omega)) \leq \sum_{\substack{\omega \in B: \\ P(\omega) < Q(\omega)}} (Q(\omega) - P(\omega)) \leq \sum_{\substack{\omega \in \Omega: \\ P(\omega) < Q(\omega)}} (Q(\omega) - P(\omega)) = d_{TV}(P, Q) \tag{52}$$

Combining:

$$2\, d_{TV}(P(\cdot \mid B), Q(\cdot \mid B)) \leq \frac{2 \cdot d_{TV}(P, Q)}{P(B)} \tag{53}$$

Since $P(B) = \min\{P(B), Q(B)\}$ under our assumption, dividing by 2 gives the result. $\square$

F.5 PROOF OF THEOREM 15

Step 1: Express interventional distributions via adjustment. By the adjustment formula:

$$\Phi_x(P) = \sum_z P(\cdot \mid X = x, \mathbf{Z} = z) \cdot P(\mathbf{Z} = z) \tag{54}$$

Step 2: Apply triangle inequality for TV distance of mixtures. For mixtures $\mu = \sum_z \alpha_z \mu_z$ and $\nu = \sum_z \beta_z \nu_z$ with $\sum_z \alpha_z = \sum_z \beta_z = 1$:

$$d_{TV}(\mu, \nu) \leq d_{TV}(\alpha, \beta) + \sum_z \beta_z \cdot d_{TV}(\mu_z, \nu_z) \tag{55}$$

Applying with $\alpha_z = P(\mathbf{Z}{=}z)$, $\beta_z = Q(\mathbf{Z}{=}z)$, $\mu_z = P(\cdot \mid x, z)$, $\nu_z = Q(\cdot \mid x, z)$:

$$d_{TV}(\Phi_x(P), \Phi_x(Q)) \leq d_{TV}(P_{\mathbf{Z}}, Q_{\mathbf{Z}}) + \sum_z Q(\mathbf{Z}{=}z) \cdot d_{TV}(P(\cdot \mid x, z), Q(\cdot \mid x, z)) \tag{56}$$

Step 3: Bound each term. Term 1: $d_{TV}(P_{\mathbf{Z}}, Q_{\mathbf{Z}}) \leq d_{TV}(P, Q)$ (marginals have smaller TV distance). Term 2: By Lemma 14 with $B = \{X{=}x, \mathbf{Z}{=}z\}$:

$$d_{TV}(P(\cdot \mid x, z), Q(\cdot \mid x, z)) \leq \frac{d_{TV}(P, Q)}{\min\{P(x, z), Q(x, z)\}} \leq \frac{d_{TV}(P, Q)}{\alpha_{\min}} \tag{57}$$

Step 4: Combine.

$$d_{TV}(\Phi_x(P), \Phi_x(Q)) \leq d_{TV}(P, Q) + \sum_z Q(z) \cdot \frac{d_{TV}(P, Q)}{\alpha_{\min}} = \left(1 + \frac{1}{\alpha_{\min}}\right) d_{TV}(P, Q) \tag{58}$$

Taking $\min$ with 1:

$$W(\mathcal{P}, x) \leq \min\left\{1, \left(1 + \frac{1}{\alpha_{\min}}\right) \operatorname{diam}_{TV}(\mathcal{P})\right\} \qquad \square \tag{59}$$

F.6 PROOF OF THEOREM 24

Lower bound. Consider any $\pi \in \Delta(\mathcal{A})$. Define $p_0 = \pi(a_0)$ and $p_1 = \pi(a_1)$.

Under P_0: $R(\pi, P_0) = U_{P_0}(a_0) - U_{P_0}(\pi) \geq (1 - p_0)g$ (since not playing a_0 with probability $1 - p_0$ incurs at least $(1 - p_0)g$ regret).

Under P_1: $R(\pi, P_1) = U_{P_1}(a_1) - U_{P_1}(\pi) \geq (1 - p_1)g$.

For actions $a \neq a_0, a_1$: their contribution to regret is at least g under both distributions (by condition (iii)).

Therefore: $\max\{R(\pi, P_0), R(\pi, P_1)\} \geq \max\{(1 - p_0)g, (1 - p_1)g\}$. This is minimized when $p_0 = p_1 = 1/2$ (and no mass on other actions), giving $\max\{R(\pi, P_0), R(\pi, P_1)\} \geq g/2$.

Upper bound (achievability). The strategy $\pi^* = \frac{1}{2}\delta_{a_0} + \frac{1}{2}\delta_{a_1}$ achieves:

Under P_0: $R(\pi^*, P_0) = U_{P_0}(a_0) - \frac{1}{2}U_{P_0}(a_0) - \frac{1}{2}U_{P_0}(a_1) = \frac{1}{2}(U_{P_0}(a_0) - U_{P_0}(a_1)) = g/2$.

Under P_1: $R(\pi^*, P_1) = g/2$ (symmetric).

So $\max_P R(\pi^*, P) = g/2$, matching the lower bound. $\square$

F.7 EXACT MINIMAX REGRET COMPUTATION

Proposition 31 (Exact Minimax Regret Computation)**.** *Let $\mathcal{A}$ be finite. Suppose the credal set $\mathcal{P}$ is specified directly over interventional distributions $\{P(Y \mid \mathrm{do}(a))\}_{a\in\mathcal{A}}$ and is polytopic. For any fixed action $a \in \mathcal{A}$, define*

$$r(a) := \max_{P\in\mathcal{P}} \left[\max_{a'\in\mathcal{A}} U_P(a') - U_P(a)\right] \tag{60}$$

Then $r(a) = \max_{a'\in\mathcal{A}} t(a,a')$ where $t(a,a') := \max_{P\in\mathcal{P}}(U_P(a') - U_P(a))$. For each fixed pair (a,a'), $t(a,a')$ is the optimal value of a linear program (linear objective over the polytope $\mathcal{P}$). Hence $r(a)$ is computable by solving $|\mathcal{A}|$ LPs, and the pure minimax-regret action is $a^* = \arg\min_{a\in\mathcal{A}} r(a)$.

If instead the credal set is over observational distributions and utilities require adjustment (ratios), then $U_P(a)$ becomes nonlinear in joint-probability decision variables and this LP reduction does not apply directly.

G NOTATION SUMMARY

Symbol	Definition	First Use
$\mathcal{P}$	Credal set (convex set of distributions)	Def. 9
Δ	Credal set diameter: $\mathrm{diam}_{TV}(\mathcal{P})$	—
P^*	True (unknown) distribution, $P^* \in \mathcal{P}$	Sec. 3
$\mathcal{G}$	Causal graph	Sec. 3
$\mathrm{do}(X\!=\!x)$	Intervention setting X to x	—
$\mathbf{Z}$	Adjustment set	Sec. 3
$\theta(P)$	Causal effect functional: $\mathbb{E}_P[Y \mid \mathrm{do}(X\!=\!x)]$	Def. **??**
π_0	Marginal positivity of hard pair	Prop. 2
$\alpha_{\min}$	Joint positivity: $\inf_P \min_z P(X\!=\!x, Z\!=\!z)$	Def. 12
$\beta_{\min}$	Conditional positivity (propensity): $\inf_P \min_z P(X\!=\!x \mid Z\!=\!z)$	Def. 4
$\kappa(\mathcal{P})$	Effective overlap: $\sup_P \mathbb{E}_P[1/P(X\!=\!x \mid Z)]$	Def. 4
$W(\mathcal{P}, x)$	Robust identification width	Def. 10
Φ_x	Interventional mapping: $P \mapsto P(Y \mid \mathrm{do}(X\!=\!x))$	Def. 9
$R(a, P)$	Regret of action a under distribution P	Def. 21
g	Action-separation gap	Def. 22

Key distinctions: π_0 (Proposition 2) denotes the marginal treatment probability of a specific hard pair; $\alpha_{\min}$ (Theorem 15) is the joint positivity used for Lipschitz bounds; $\beta_{\min}$ (Definition 4) is conditional positivity (propensity) for the adjustment case. These are different positivity notions serving different purposes.