Data-Driven Information-Theoretic Causal Bounds under Unmeasured Confounding

Yonghan Jung[1] and Bogyeong Kang[2]

[1] University of Illinois Urbana-Champaign, yonghan@illinois.edu

[2] Independent Researcher, bgyeong.kang@gmail.com

January 23, 2026

We develop a data-driven information-theoretic framework for the sharp partial identification of causal effects under unmeasured confounding. Existing approaches often rely on restrictive assumptions, such as bounded or discrete outcomes, require external inputs (e.g., instrumental variables, proxies, or user-specified sensitivity parameters), necessitate full structural causal model specifications, or focus solely on population-level averages while neglecting covariate-conditional treatment effects. We overcome all four limitations simultaneously by establishing novel information-theoretic, data-driven divergence bounds. Our key theoretical contribution establishes that the f-divergence between the observational distribution $P(Y \mid A = a, X = x)$ and the interventional distribution $P(Y \mid \mathrm{do}(A = a), X = x)$ is upper bounded by a function of the propensity score alone. This result enables sharp partial identification of conditional causal effects directly from observational data, without requiring external sensitivity parameters, auxiliary variables, full structural specifications, or outcome boundedness assumptions. For practical implementation, we develop a semiparametric estimator satisfying Neyman-orthogonality (Chernozhukov et al., 2018), which ensures $\sqrt{n}$-consistent inference even when nuisance functions are estimated via flexible machine learning methods. Simulation studies and real-world data applications demonstrate that our framework provides tight and valid causal bounds across a wide range of data-generating processes.

Contents

1. Introduction

Causal effect identification aims to characterize interventional quantities, such as $\Pr(Y = y \mid \mathrm{do}(A = a), X = x)$, as functionals of the observational distribution $P(X, A, Y)$. In the presence of unmeasured confounders U, as depicted in Fig. 1a, point identification is generally impossible without auxiliary variables or structural restrictions. In such settings, *partial identification* seeks to recover bounds that provably contain the true causal quantity. However, as described in literature review in Sec. 1.1, most existing methods suffer from one or more of the following fundamental limitations:

(Lim-1) Bounded outcomes: Restricting outcomes to bounded or discrete supports (e.g., $Y \in [0, 1]$).

(Lim-2) Externality of parameters: Requiring auxiliary inputs—such as instrumental variables, proxies, or sensitivity parameters—to quantify confounding strength.

(Lim-3) Full SCM specification: Necessitating the specification of the entire structural causal model (SCM) (Pearl, 2000), which is computationally intensive and prone to error propagation.

(Lim-4) Neglect of heterogeneity: Focusing on population-level averages while neglecting covariate-conditional treatment effects.

To universally address these limitations, we develop an information-theoretic framework that provides (i) data-driven upper bounds on statistical divergences between observational and interventional distributions, and (ii) sharp partial identification of conditional causal effects $\mathbb{E}[Y \mid \mathrm{do}(A = a), X = x]$. Our framework accommodates *unbounded* continuous outcomes without requiring full structural modeling or external inputs. The core mechanism involves deriving *data-driven* upper bounds on statistical divergences (e.g., f-divergence (Csiszár, 1967)) between the interventional law $Q_{a,x}$ and the observational law $P_{a,x}$, and then translating these into sharp causal intervals. Specifically, we make three main contributions:

(i) We show that f-divergences (Csiszár, 1967) between $P_{a,x}$ and $Q_{a,x}$ are upper bounded by a function of the propensity score $e_a(x)$.

(ii) We leverage these bounds to obtain sharp intervals for arbitrary expectations of the form $\theta(a, x) \triangleq \mathbb{E}_{Q_{a,x}}[\varphi(Y)]$ for user-specified functions φ without imposing outcome boundedness or support restrictions.

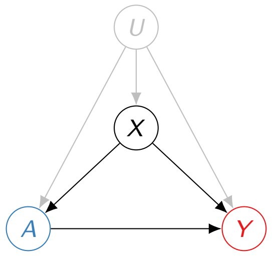
(iii) We develop a semiparametric estimator that satisfies Neyman-orthogonality (Chernozhukov et al., 2018) ensuring robust inference even when nuisance components are estimated via high-dimensional machine learning models.

Together, these results provide a principled path to *data-driven* partial identification of conditional causal effects under unmeasured confounding.

1.1. Related Work

We organize existing work on partial identification based on which of the limitations (Lim-(1,2,3,4)) they retain or address.

Bounded/discrete outcomes (Lim-1). Early work imposed restrictions requiring outcomes to be bounded or discrete. For example, Manski (1990) derived nonparametric bounds using



(a)

Method	Unbounded Outcome	No Aux Input	No Full SCM	Cond. Effect
LP / Discrete	✗	✓	✓	✗
Additional Vars. (IV, Proxy)	✗	✗	✓	✓
Sensitivity	✓	✗	✓	✓
Full SCM	✓	✓	✗	✓
Ours	✓	✓	✓	✓

(b)

Figure 1: **(a)** Causal diagram with unmeasured confounding. **(b)** Systematic comparison of our method against existing literature (detailed in Sec. 1.1).

the extreme values outcomes can attain. Linear-programming (LP)–based approaches (e.g., Balke and Pearl (1994), Tian and Pearl (2000)) yield sharp bounds with discrete variables. Sachs et al. (2023) and Shridharan and Iyengar (2023) have extended these LP-based bounds to general graphical settings but remain restricted to discrete outcomes. Zhang and Bareinboim (2021) extended these LP ideas to continuous outcomes, but still rely on bounded-support assumptions (e.g., $Y \in [0, 1]$). These methods avoid auxiliary inputs (addressing Lim-2) but fail to accommodate unbounded outcomes (Lim-1) or provide conditional effect bounds (Lim-4).

Auxiliary inputs (Lim-2). Another line of work leverages auxiliary inputs. While auxiliary-variable methods can yield sharp bounds, most methods still assume bounded outcomes (Lim-1); and valid auxiliary inputs are often not available in practice or not identifiable from data.

- *Instrumental variables.* Balke and Pearl (1997) provide tight nonparametric bounds on average treatment effects by leveraging instrumental variables, assuming bounded binary outcomes. Kitagawa (2021) extends this framework to continuous outcomes while maintaining bounded support assumptions (see Swanson et al. (2018) for a comprehensive survey). Recently, Levis et al. (2025) develop covariate-assisted IV bounds to target conditional treatment effects (addressing Lim-4), but also under bounded outcome assumptions (Lim-1).
- *Additional assumptions or variables.* Ghassami et al. (2023) leverage proxy variables of hidden confounders to provide bounds on average effects, again requiring bounded outcomes (Lim-1). Lee (2009) and Semenova (2025) avoid bounded outcomes for sharp bounds on the average effect, but rely on structural assumptions about the selection mechanism.
- *Sensitivity analysis.* Sensitivity analysis introduces user-specified parameters to quantify confounding strength (e.g., Rosenbaum (1987), Tan (2006), Yadlowsky et al. (2022), Jin et al. (2022), Dorn and Guo (2023), Oprescu et al. (2023)). Unlike IV and proxy methods, modern sensitivity approaches can accommodate unbounded outcomes (addressing Lim-1). Among these, Jin et al. (2022) are most closely related to our approach, as they use an f-divergence-based sensitivity model to constrain divergences between observational and interventional distributions. Oprescu et al. (2023) extend sensitivity analysis to bound conditional effects (addressing Lim-4). However, all sensitivity methods require external sensitivity parameters (Lim-2) that are not identifiable from observational data alone.

Full SCM-modeling approaches (Lim-3). Another approach leverages machine-learning methods to learn entire SCMs consistent with observational data (e.g., Hu et al. (2021), Balazadeh Meresht et al. (2022), Padh et al. (2023), Xia et al. (2022), Tan et al. (2024)). These approaches find the SCMs that maximize/minimize the target causal effect subject to observations, using flexible neural architectures to model structural functions. In principle, such methods can accommodate unbounded outcomes and target conditional effects (addressing Lim-(1,4)). However, they require estimating the entire SCM (Lim-3), which is computationally intensive and sensitive to misspecification in high-dimensional structural components.

Our novelty. Existing methods each resolve some limitations; however, no existing approach achieves all four limitations (Lim-1–Lim-4) universally. In contrast, our work simultaneously addresses all four limitations by developing bounds that (Lim-1) accommodate unbounded continuous outcomes without support restrictions; (Lim-2) require no auxiliary variables or sensitivity parameters; (Lim-3) avoid full SCM modeling; and (Lim-4) provide bounds for conditional effects $\mathbb{E}[Y \mid \mathrm{do}(A=a), X=x]$ beyond the population-level average. We compare our work with representative existing methods in Fig. 1b.

2. Problem Setup & Preliminaries

Consider a treatment $A \in \{0,1\}$, a covariates vector $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, and an outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. We consider the structural causal model (SCM) framework (Pearl, 2000) as the data-generating process (DGP) for (X, A, Y):

$$U \leftarrow f_U(\epsilon_U), \;\; X \leftarrow f_X(U, \epsilon_X), \;\; A \leftarrow f_A(X, U, \epsilon_A), \;\; Y \leftarrow f_Y(X, A, U, \epsilon_Y), \tag{1}$$

where U represents unmeasured confounding, $f_{(\cdot)}$ are unknown structural functions, and $(\epsilon_U, \epsilon_X, \epsilon_A, \epsilon_Y)$ are mutually independent exogenous noise variables. The causal diagram induced by this SCM is depicted in Fig. 1a.

The operation $\mathrm{do}(A=a)$ denotes an intervention that replaces f_A with a constant $a \in \{0,1\}$, while keeping the other structural equations invariant. For each $(a,x) \in \{0,1\} \times \mathcal{X}$, we define the following conditional probability laws on $\mathcal{Y}$:

- **Observational Law**: $P_{a,x} \equiv P(Y \mid A=a, X=x)$, which is identifiable from data.
- **Interventional Law**: $Q_{a,x} \equiv P(Y \mid \mathrm{do}(A=a), X=x)$, our target of interest.

Under unmeasured confounding (i.e., when f_A and f_Y share U as a common hidden parent), the interventional law $Q_{a,x}$ is unidentifiable from the observational law $P_{a,x}$. Consequently, any causal functional $\theta = \mathbb{E}_{Q_{a,x}}[\varphi(Y)]$ for some user-specified φ (e.g., the identity for ATE/CATE) is also unidentifiable.

f-Divergence. To characterize the "distance" between the identifiable $P_{a,x}$ and the unidentifiable $Q_{a,x}$, we use f-divergences (Ali and Silvey, 1966; Csiszár, 1967).

Definition 1 (f-Divergence). Let P and Q be probability measures on $(\mathcal{Y}, \mathcal{F})$ such that $P \ll Q$. For a convex function $f : [0,\infty) \to \mathbb{R}$ with $f(1)=0$, the f-divergence of P from Q is

$$D_f(P \| Q) \triangleq \int_{\mathcal{Y}} f\left(\frac{dP}{dQ}\right) dQ. \tag{2}$$

Common specializations of f-divergence are as follows. Let p and q be the Radon–Nikodym derivatives of P and Q with respect to a common dominating measure μ (e.g., Lebesgue or counting measure).

- **Kullback-Leibler (KL).** $f(t) \triangleq t \log t$ with $f(0) = 0$. Then,

$$D_{\mathrm{KL}}(P\|Q) = \int_{\mathcal{Y}} \log\left(\frac{dP}{dQ}\right) dP = \int_{\mathcal{Y}} p(y) \log\left(\frac{p(y)}{q(y)}\right) d\mu(y). \tag{3}$$

- **Hellinger distance.** $f(t) \triangleq \frac{1}{2}(\sqrt{t} - 1)^2$ with $f(0) = 1/2$.

$$D_{\mathrm{H}}(P\|Q) = \frac{1}{2} \int_{\mathcal{Y}} \left(\sqrt{\frac{dP}{dQ}} - 1\right)^2 dQ = 1 - \int_{\mathcal{Y}} \sqrt{p(y)q(y)} d\mu(y). \tag{4}$$

- **χ^2-divergence.** $f(t) \triangleq \frac{1}{2}(t - 1)^2$ with $f(0) = 1/2$.

$$D_{\chi^2}(P\|Q) = \frac{1}{2} \int_{\mathcal{Y}} \left(\frac{dP}{dQ} - 1\right)^2 dQ = \frac{1}{2} \int_{\mathcal{Y}} \frac{(p(y) - q(y))^2}{q(y)} d\mu(y). \tag{5}$$

- **Total variation (TV).** $f(t) \triangleq \frac{1}{2}|t - 1|$ with $f(0) = 1/2$.

$$D_{\mathrm{TV}}(P\|Q) = \frac{1}{2} \int_{\mathcal{Y}} |p(y) - q(y)| d\mu(y) = \sup_{B \in \mathcal{F}} |P(B) - Q(B)|. \tag{6}$$

- **Jensen-Shannon.** $f(t) \triangleq \frac{1}{2}(t \log t - (t + 1) \log(\frac{t+1}{2}))$ with $f(0) = \frac{1}{2} \log 2$. Let $M \triangleq \frac{P+Q}{2}$.

$$D_{\mathrm{JS}}(P\|Q) = \frac{1}{2} D_{\mathrm{KL}}(P\|M) + \frac{1}{2} D_{\mathrm{KL}}(Q\|M). \tag{7}$$

Integral Probability Metrics (IPMs) & Maximum Mean Discrepancy (MMD). Beyond the f-divergence, the integral probability metric (IPM; Müller 1997) and maximum mean discrepancy (MMD; Gretton et al. 2012) are commonly used. Let $\Phi \triangleq \{\varphi : \mathcal{Y} \mapsto [0, 1]\}$ be a class of measurable functions. Then, each measures are defined as follows.

Definition 2 (**Integral Probability Metric (IPM)** (Müller, 1997))**.** Let P and Q be probability measures on a measurable space $(\mathcal{Y}, \mathcal{F})$. Let Φ be a class of measurable real-valued functions on $\mathcal{Y}$. The integral probability metric (IPM) is

$$D_{\mathrm{IPM},\Phi}(P\|Q) \triangleq \sup_{\varphi \in \Phi} \left|\mathbb{E}_P[\varphi(Y)] - \mathbb{E}_Q[\varphi(Y)]\right|. \tag{8}$$

Definition 3 (**Maximum Mean Discrepancy (MMD)** (Gretton et al., 2012))**.** Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space (RKHS) associated with a positive-definite kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The maximum mean discrepancy (MMD) is

$$D_{\mathrm{MMD},k}(P\|Q) \triangleq \sup_{\|h\|_{\mathcal{H}_k} \leq 1} \left|\mathbb{E}_P[h(Y)] - \mathbb{E}_Q[h(Y)]\right|. \tag{9}$$

When the function class Φ is sufficiently rich (e.g., all bounded continuous functions), the IPM fully characterizes distributional differences. Similarly, when the kernel k is characteristic, the MMD fully characterizes distributional differences. In both cases, $D_{\mathrm{IPM},\Phi}(P\|Q) = 0$ (or $D_{\mathrm{MMD},k}(P\|Q) = 0$) if and only if $P = Q$. These quantities can be viewed as special cases of distributional discrepancies defined through restricted function classes, and serve as limiting or simplified alternatives to the f-divergence–based bounds considered in the main text.

Problem Statement. Under the assumed DGP in Fig. 1a and Eq. (1), the interventional law $Q_{a,x}$ and a target causal effect in the form of $\mathbb{E}_{Q_{a,x}}[\varphi(Y)]$ (where φ is a user-specified and potentially continuous and unbounded function) are generally not identifiable from the observational law $P_{a,x}$ due to unmeasured confounding. To address, (1) we derive the upper limit of the f-divergence of the observational law $P_{a,x}$ from the interventional law $Q_{a,x}$; i.e., $D_f(P_{a,x}\|Q_{a,x})$; (2) we translate the upper limit of the f-divergence into the sharp interval for causal effects. Throughout the paper, we assume the following:

Assumption 1. For all $a, x \in \mathcal{A} \times \mathcal{X}$,

1. **Positivity**: $e_a(x) \triangleq \Pr(A = a \mid X = x) \in [c, 1-c]$ for some constant $0 < c < 1/2$.
2. **Mutual absolute continuity**: $P_{a,x} \ll Q_{a,x}$ and $Q_{a,x} \ll P_{a,x}$.
3. **Regularity of f**: For the generator function f in the f-divergence, $f(0) < \infty$.

3. Divergence Bounds between Observational and Interventional Distributions

We now derive a data-driven upper bound on the f-divergence between observational and interventional distributions. Our main result is the following:

Theorem 1 (f-Divergence Bound). For any $a \in \mathcal{A}$ and $x \in \mathcal{X}$ such that $P(a \mid x) > 0$,

$$D_f(P_{a,x}\|Q_{a,x}) \leq B_f(e_a(x)), \tag{10}$$

where

$$B_f(e_a(x)) \triangleq e_a(x) f\left(\frac{1}{e_a(x)}\right) + (1 - e_a(x)) f(0) \tag{11}$$

Theorem 1 establishes that the f-divergence $D_f(P_{a,x}\|Q_{a,x})$ is upper bounded by $B_f(e_a(x))$, a function of the propensity score that is directly computable from observational data. Notably, $B_f(e_a(x)) \to 0$ as $e_a(x) \to 1$, since f is continuous (by convexity) and satisfies $f(1) = 0$. Thus, higher propensity scores yield tighter divergence bounds.

We specialize Thm. 1 to standard divergences:

Corollary T1.1. For any $a \in \mathcal{A}$ and $x \in \mathcal{X}$ such that $P(a \mid x) > 0$,

- **KL**: $f(t) \triangleq t \log t$ (with $f(0) = 0$),

$$D_{\mathrm{KL}}(P_{a,x}\|Q_{a,x}) \leq -\log e_a(x). \tag{12}$$

- **Hellinger**: $f(t) \triangleq \frac{1}{2}(\sqrt{t}-1)^2$ (with $f(0)=1/2$),

$$D_{\mathrm{H}}(P_{a,x}\|Q_{a,x}) \le 1-\sqrt{e_a(x)}. \tag{13}$$

- χ^2**-divergence**: $f(t) \triangleq \frac{1}{2}(t-1)^2$ (with $f(0)=1/2$),

$$D_{\chi^2}(P_{a,x}\|Q_{a,x}) \le \frac{1-e_a(x)}{2e_a(x)}. \tag{14}$$

- **Total variation**: $f(t) \triangleq \frac{1}{2}|t-1|$ (with $f(0)=\frac{1}{2}$),

$$D_{\mathrm{TV}}(P_{a,x}\|Q_{a,x}) \le 1-e_a(x). \tag{15}$$

- **Jensen-Shannon**: $f(t) \triangleq f_{\mathrm{JS}}(t) \triangleq \frac{1}{2}\left(t\log t-(t+1)\log\left(\frac{t+1}{2}\right)\right)$ (with $f_{\mathrm{JS}}(0) = \frac{1}{2}\log 2$)

$$D_{\mathrm{JS}}(P_{a,x}\|Q_{a,x}) \le B_{f_{\mathrm{JS}}}(e_a(x)) = \frac{1}{2}\log\left(\frac{4e_a(x)^{e_a(x)}}{(1+e_a(x))^{1+e_a(x)}}\right). \tag{16}$$

Bounds extend to stochastic policies as follows:

Corollary T1.2. For any stochastic policy $\pi(a \mid x)$,

$$D_f(P_\pi\|Q_\pi) \triangleq \mathbb{E}_X\left[\sum_{a\in\mathcal{A}}\pi(a\mid X)D_f(P_{a,X}\|Q_{a,X})\right] \le \mathbb{E}_X\left[\sum_{a\in\mathcal{A}}\pi(a\mid X)B_f(e_a(X))\right].$$

Choosing $\pi(a \mid x) = e_a(x)$ yields the global divergence bound:

$$D_f(P_{A,X}\,\|\,Q_{A,X}) = \mathbb{E}_X\left[\sum_{a\in\mathcal{A}}e_a(X)\,D_f(P_{a,X}\,\|\,Q_{a,X})\right] \;\le\; \mathbb{E}_X\left[\sum_{a\in\mathcal{A}}e_a(X)\,B_f(e_a(X))\right].$$

We derive bounds on the maximum mean discrepancy (MMD; Gretton et al. 2012), and the integral probability metric (IPM; Müller 1997).

Corollary T1.3 (IPM and MMD Bounds). Let $\Phi \triangleq \{\varphi : \mathcal{Y} \mapsto [0,1]\}$ be a class of measurable functions. Let $D_{\mathrm{IPM},\mathcal{F}}(P\|Q)$ be the IPM over a function class $\mathcal{F} \triangleq \{f : \|f\|_\infty < C\}$. Let $D_{\mathrm{MMD},\mathtt{k}}(P\|Q)$ be the MMD associated with an RKHS with a kernel $\mathtt{k}$ such that $\mathtt{k}(\cdot,\cdot) < K$. Then,

$$\textbf{(IPM)}\;\; D_{\mathrm{IPM},\mathcal{F}_C}(P_{a,x}\,\|\,Q_{a,x}) \le 2C\,\min\left\{1-e_a(x),\,\sqrt{-\tfrac{1}{2}\log e_a(x)}\right\},$$
$$\textbf{(MMD)}\;\; D_{\mathrm{MMD},\mathtt{k}}(P_{a,x}\,\|\,Q_{a,x}) \le 2\sqrt{K}\,\min\left\{1-e_a(x),\,\sqrt{-\tfrac{1}{2}\log e_a(x)}\right\}.$$

All results above extend to the marginal case (without covariates) by setting $X = \emptyset$ and replacing $e_a(x)$ with the marginal propensity score $e_a \triangleq \Pr(A=a)$. This yields bounds on the divergence between the marginal interventional law $Q_a \triangleq P(Y \mid \mathrm{do}(A=a))$ and the marginal observational law $P_a \triangleq P(Y \mid A=a)$.

3.1. Specialization for Exponential Family

Here, we derive a closed-form f-divergence when $P_{a,x}$, $Q_{a,x}$ are within the exponential family (e.g., Bernoulli, Gaussian, Poisson, exponential, etc.) to exemplify the mechanism of Thm. 1.

Corollary 4 (Exponential Family). Suppose $P_{a,x}$ and $Q_{a,x}$ are distributions from a common exponential family:

$$P_{a,x}(y) \triangleq \exp\left(\theta_p^\top T(y) - A(\theta_p)\right)h(y), \tag{17}$$
$$Q_{a,x}(y) \triangleq \exp\left(\theta_q^\top T(y) - A(\theta_q)\right)h(y), \tag{18}$$

where θ_p, θ_q are natural parameters, $T(y)$ is the sufficient statistics, $A(\theta)$ is the log-partition function (log normalizer), and $h(y)$ is the base measure density. Define $\Delta \triangleq \theta_p - \theta_q$ and $\Delta_A \triangleq A(\theta_p) - A(\theta_q)$.

$$D_f(P_{a,x} \| Q_{a,x}) = \mathbb{E}_{Q_{a,x}}\left[f\left(\exp\left(\Delta^\intercal T(Y) - \Delta_A\right)\right)\right]. \tag{19}$$

Bernoulli Distribution Suppose $Y \in \{0,1\}$ and both $P_{a,x}$ and $Q_{a,x}$ are Bernoulli distributions with success probabilities p and q, respectively. The Bernoulli distribution belongs to the exponential family with sufficient statistic $T(y) = y$ and natural parameter $\theta = \log \frac{p}{1-p}$.

The Radon–Nikodym derivative is given by

$$\frac{dP_{a,x}}{dQ_{a,x}}(y) = \exp\left(y \log \frac{p(1-q)}{q(1-p)} + \log \frac{1-p}{1-q}\right). \tag{20}$$

Consequently, the f-divergence admits the representation

$$D_f(P_{a,x} \| Q_{a,x}) = \mathbb{E}_{Q_{a,x}}\left[f\left(\exp\left(Y \log \frac{p(1-q)}{q(1-p)} + \log \frac{1-p}{1-q}\right)\right)\right]. \tag{21}$$

Gaussian Distribution Suppose $Y \in \mathbb{R}^d$ and both $P_{a,x}$ and $Q_{a,x}$ are Gaussian distributions with means μ_p, μ_q and covariance matrices Σ_p and Σ_q, respectively. In this case, the Gaussian distribution forms an exponential family with sufficient statistic $T(y) = (y, yy^\top)$.

The Radon–Nikodym derivative is given by

$$\frac{dP_{a,x}}{dQ_{a,x}}(y) = \frac{|\Sigma_q|^{1/2}}{|\Sigma_p|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu_p)^\top \Sigma_p^{-1}(y-\mu_p) + \frac{1}{2}(y-\mu_q)^\top \Sigma_q^{-1}(y-\mu_q)\right). \tag{22}$$

Accordingly, using (22), the f-divergence can be written as

$$D_f(P_{a,x} \| Q_{a,x}) = \mathbb{E}_{Q_{a,x}}\left[f\left(\frac{dP_{a,x}}{dQ_{a,x}}(Y)\right)\right]. \tag{23}$$

Poisson Distribution Suppose $Y \in \{0, 1, 2, \dots\}$ and both $P_{a,x}$ and $Q_{a,x}$ are Poisson distributions with rate parameters λ_p and λ_q, respectively. The Poisson distribution belongs to the exponential family with sufficient statistic $T(y) = y$ and natural parameter $\theta = \log \lambda$.

The Radon–Nikodym derivative takes the form

$$\frac{dP_{a,x}}{dQ_{a,x}}(y) = \exp\left(y \log \frac{\lambda_p}{\lambda_q} - (\lambda_p - \lambda_q)\right). \tag{24}$$

The corresponding f-divergence is therefore

$$D_f(P_{a,x} \| Q_{a,x}) = \mathbb{E}_{Q_{a,x}}\left[f\left(\exp\left(Y \log \frac{\lambda_p}{\lambda_q} - (\lambda_p - \lambda_q)\right)\right)\right]. \tag{25}$$

Exponential Distribution Suppose $Y \geq 0$ and both $P_{a,x}$ and $Q_{a,x}$ follow exponential distributions with rate parameters λ_p and λ_q, respectively. The exponential distribution is an exponential family with sufficient statistic $T(y) = y$ and natural parameter $\theta = -\lambda$.

For $y \geq 0$, the Radon–Nikodym derivative is given by

$$\frac{dP_{a,x}}{dQ_{a,x}}(y) = \frac{\lambda_p}{\lambda_q} \exp\left(-(\lambda_p - \lambda_q)y\right). \tag{26}$$

Accordingly, the f-divergence can be expressed as

$$D_f(P_{a,x} \| Q_{a,x}) = \mathbb{E}_{Q_{a,x}}\left[f\left(\frac{\lambda_p}{\lambda_q} \exp\left(-(\lambda_p - \lambda_q)Y\right)\right)\right]. \tag{27}$$

4. A Distributionally Robust Formulation of Causal Bounds

In this section, we leverage the upper bounds on statistical divergence derived in Section 3 to construct bounds on the target causal effect $\theta(a,x) \triangleq \mathbb{E}_{Q_{a,x}}[\varphi(Y)]$, where $\varphi(Y)$ is an arbitrary measurable function with finite first and second moments. This framework encompasses diverse causal quantities: setting $\varphi(Y) \triangleq \mathbf{1}(Y \leq t)$ yields the cumulative distribution function $Q_{a,x}(Y \leq t)$, while choosing $\varphi(Y) \triangleq \ell(Y;\theta)$ (a loss function for θ) yields the risk function over $Q_{a,x}$. Crucially, we impose no restrictions requiring φ to be discrete or bounded.

Using the divergence bound $D_f(P_{a,x} \| Q_{a,x}) \leq B_f(e_a(x))$ from Thm. 1, we define the f-divergence-based *ambiguity set*, which is a collection of distributions over Y within the $B_f(e_a(x))$ radius around the observational law $P_{a,x}$:

$$\textbf{(Ambiguity set)}\quad \mathcal{Q}_f(a,x;P_{a,x}) \triangleq \left\{ \begin{aligned} Q_{a,x} \in \mathcal{P}_{a,x}(\mathcal{Y}) \quad : \quad & D_f(P_{a,x} \,\|\, Q_{a,x}) \leq B_f(e_a(x)), \\ & P_{a,x} \ll Q_{a,x} \end{aligned} \right\}, \tag{28}$$

where $\mathcal{P}_{a,x}(Y)$ is a collection of probability laws given $A = a$ and $X = x$. The target causal effect $\mathbb{E}_{Q_{a,x}}[\varphi(Y)]$ is bounded by expectations over the extremal distributions in this ambiguity set:

$$\textbf{(Bounds)} \qquad \underbrace{\theta_{\mathrm{lo}}(a,x)}_{\inf_{Q \in \mathcal{Q}_f(a,x)} \mathbb{E}_Q[\varphi(Y)]} \leq \underbrace{\theta(a,x)}_{\mathbb{E}_{Q_{a,x}}[\varphi(Y)]} \leq \underbrace{\theta_{\mathrm{up}}(a,x)}_{\sup_{Q \in \mathcal{Q}_f(a,x)} \mathbb{E}_Q[\varphi(Y)]} \tag{29}$$

The lower and upper bounds are symmetric: by Proposition 1 below, the lower bound can be obtained from the upper bound by negating the function φ. Therefore, we focus on deriving the upper bound $\theta_{\mathrm{up}}(a,x)$ without loss of generality.

Proposition 1 (Lower bound as a subproblem of upper bound). Let

$$\theta_{\text{lo}}(a,x;\varphi) \triangleq \inf_{Q\in\mathcal{Q}_f(a,x)} \mathbb{E}_Q[\varphi(Y)], \qquad \theta_{\text{up}}(a,x;\varphi) \triangleq \sup_{Q\in\mathcal{Q}_f(a,x)} \mathbb{E}_Q[\varphi(Y)]. \tag{30}$$

Then,

$$\theta_{\text{lo}}(a,x;\varphi) = -\theta_{\text{up}}(a,x;-\varphi). \tag{31}$$

By Proposition 1, it suffices to compute $\theta_{\text{up}}(a,x)$. However, computing $\theta_{\text{up}}(a,x)$ directly from Eq. (29) is intractable, as it requires optimizing over the infinite-dimensional space of all probability measures in $\mathcal{Q}_f(a,x)$. To overcome this computational barrier, we reformulate the problem using convex duality:

Theorem 2 (Primal and Dual Formulations). Let $s(Y) \triangleq \frac{dQ_{a,x}}{dP_{a,x}}(Y)$ denote the likelihood ratio, $g_s(Y) \triangleq s(Y)\cdot f(1/s(Y))$, and $\eta_f(a,x) \triangleq B_f(e_a(x))$. The upper bound $\theta_{\text{up}}(a,x)$ admits the following equivalent representations:

$$\theta_{\text{up}}(a,x) = \sup_{s>0}\left\{\mathbb{E}_{P_{a,x}}[s(Y)\varphi(Y)] \text{ s.t. } \mathbb{E}_{P_{a,x}}[s(Y)]=1,\ \mathbb{E}_{P_{a,x}}[g_s(Y)] \le \eta_f(a,x)\right\} \tag{32}$$

$$= \inf_{\lambda>0,u\in\mathbb{R}}\{\lambda\eta_f(a,x) + u + \lambda\mathbb{E}_{P_{a,x}}\left[g^*\left(\tfrac{\varphi(Y)-u}{\lambda}\right)\right]\}, \tag{33}$$

where $g^*(t) \triangleq \sup_{s>0}\{s\,t - g(s)\}$ is the convex conjugate (also known as the Legendre–Fenchel conjugate or c-transform) of g.

The following proposition provides a general recipe for computing the convex conjugate g^*:

Proposition 2 (Convex Conjugate g^*). Let $f:(0,\infty)\to(-\infty,\infty]$ be proper, convex, and lower semi-continuous function. Define for $s>0$,

$$g(s) \triangleq sf(1/s), \quad g^*(t) \triangleq \sup_{s>0}\{st - g(s)\}. \tag{34}$$

Let $r \triangleq 1/s$. Then,

$$g^*(t) \triangleq \sup_{r>0} \frac{t - f(r)}{r}. \tag{35}$$

Moreover, if the supremum is attained at some $r^*>0$, then there exists a subgradient $a \in \partial f(r^*)$ such that

$$t = f(r^*) - r^*a, \quad \text{and} \quad g^*(t) = -a. \tag{36}$$

If f is differentiable at r^*, then $a = f'(r^*)$ and hence $g^*(t) = -f'(r^*)$.

[YJ - (Make this as more verbally accessible and easier statement for explaining Prop. 2) Prop. 2 basically conducts the change-of-variables $r \triangleq 1/s$, and, applies the optimality condition, stating that if the maximizer r^* exists, then there is a subgradient $a \in \partial f(r^*)$ such that $t = f(r^*) - r^*a$ and $g^*(t) = -a$.]

We apply Prop. 2 to standard f-divergences:

Corollary P1. Let $g(s) \triangleq sf(1/s)$ for $s > 0$. Then,

- **KL**: $g_{\mathrm{KL}}(s) = -\log s$, and

$$g^*_{\mathrm{KL}}(t) = \begin{cases} -1 - \log(-t) & \text{if } t < 0; \\ +\infty & \text{if } t \geq 0. \end{cases} \tag{37}$$

- **Hellinger**: $g_{\mathrm{H}}(s) = \frac{1}{2}(1 - 2\sqrt{s} + s)$, and

$$g^*_{\mathrm{H}}(t) = \begin{cases} \frac{t}{1-2t} & \text{if } t < 1/2; \\ +\infty & \text{if } t \geq 1/2. \end{cases} \tag{38}$$

- **Chi-square**: $g_{\chi^2}(s) = \frac{(1-s)^2}{2s}$, and

$$g^*_{\chi^2}(t) = \begin{cases} 1 - \sqrt{1-2t} & \text{if } t \leq 1/2; \\ +\infty & \text{if } t > 1/2. \end{cases} \tag{39}$$

- **TV**: $g_{\mathrm{TV}}(s) = \frac{1}{2}|1 - s|$, and

$$g^*_{\mathrm{TV}}(t) = \begin{cases} -\frac{1}{2}, & \text{if } t \leq -\frac{1}{2}, \\ t, & \text{if } -\frac{1}{2} < t \leq \frac{1}{2}, \\ +\infty, & \text{if } t > \frac{1}{2}. \end{cases} \tag{40}$$

- **Jensen-Shannon**: $g_{\mathrm{JS}}(s) = \frac{1}{2}\left(s \log s - (1+s)\log(1+s) + (1+s)\log 2\right)$, and

$$g^*_{\mathrm{JS}}(t) = \begin{cases} -\frac{1}{2}\log\left(2 - \exp(2t)\right), & \text{if } t < \frac{1}{2}\log 2, \\ +\infty, & \text{if } t \geq \frac{1}{2}\log 2. \end{cases} \tag{41}$$

All results above extend to the marginal case (without covariates) by setting $X = \emptyset$ and replacing $e_a(x)$ with the marginal propensity score $e_a \triangleq \Pr(A = a)$. This yields bounds on the marginal causal effect $\mathbb{E}_{Q_a}[Y] \triangleq \mathbb{E}[Y \mid \mathrm{do}(A = a)]$, where $Q_a \triangleq P(Y \mid \mathrm{do}(A = a))$.

5. Debiased Semiparametric Estimation of Causal Bounds

Solving the dual problem in Eq. (33) pointwise for each (a, x) is computationally intractable, as it requires estimating the conditional expectation $\mathbb{E}_{P_{a,x}}[g^*(\cdot)]$ separately for each pair of covariate $X = x$ and treatment $A = a$ at every optimization iteration. We circumvent this by amortizing the optimization as follows: we view $\lambda(a, x)$ and $u(a, x)$ as functional parameters to be learned globally. Parameterizing $\lambda(a, x) \triangleq \exp(h(a, x))$ to enforce positivity, the dual problem transforms into:

Proposition 3. Let $\eta_f(a, x) \triangleq B_f(e_a(x))$. Then,

$$\theta_{\mathrm{up}}(a, x) = \inf_{\substack{h(a,x)\in\mathbb{R} \\ u(a,x)\in\mathbb{R}}} \mathbb{E}_{P_{a,x}}\Big[\exp(h(A,X))\{\eta_f(A,X) + g^*\big(\tfrac{\varphi(Y)-u(A,X)}{\exp(h(A,X))}\big)\} + u(A,X)\Big]. \tag{42}$$

To operationalize this optimization, we define a loss function and corresponding risk function for the functional parameters h and u:

Definition 4 (Risk Function for Causal Bound). Let $V = (X, A, Y)$. Let $h_\beta, u_\gamma : \mathcal{A} \times \mathcal{X} \mapsto \mathbb{R}$ be maps parametrized by $\beta \in \mathbb{R}^{p_1}$ and $\gamma \in \mathbb{R}^{p_2}$. The risk function for causal bounds is

$$\mathcal{R}(\beta, \gamma; e) \triangleq \mathbb{E}_P[\ell(V; (\beta, \gamma), e)], \tag{43}$$

where $e \triangleq e_A(X)$ and $\eta_f \triangleq \eta_f(A, X) \triangleq B_f(e_A(X))$, and

$$\ell(V; (\beta, \gamma), e) \triangleq \exp(h_\beta(A, X))\Big\{\eta_f(A, X) + g^*\Big(\tfrac{\varphi(Y) - u_\gamma(A,X)}{\exp(h_\beta(A,X))}\Big)\Big\} + u_\gamma(A, X). \tag{44}$$

The following proposition shows that this risk minimization is equivalent to solving the pointwise dual problem:

Proposition 4 (Justification of Risk Function). Define, for each (a, x), the following loss

$$\ell(h, u; y, a, x) \triangleq \exp(h(a, x))\Big\{\eta_f(a, x) + g^*\big(\tfrac{\varphi(y) - u(a,x)}{\exp(h(a,x))}\big)\Big\} + u(a, x). \tag{45}$$

Let $\mathcal{R}(h, u) \triangleq \mathbb{E}_P[\ell(h, u; Y, A, X)]$ be a risk function. Assume $\mathcal{R}(h, u) < \infty$ for all $h, u \in \mathcal{F}$, where $\mathcal{F}$ is a function class rich enough that for any $(h_1, u_1), (h_2, u_2) \in \mathcal{F}$ and $\forall B \subset \mathcal{A} \times \mathcal{X}$, $(h', u') \triangleq (h_1, u_1)\mathbf{1}_B + (h_2, u_2)\mathbf{1}_{B^c}$ also lies in $\mathcal{F}$. Then, for any fixed $(h^\star, u^\star) \in \mathcal{F}$, the followings are equivalent:

1. $(h^\star, u^\star)$ minimizes $\mathcal{R}$ over $\mathcal{F}$.
2. $(h^\star, u^\star)$ minimizes $\mathbb{E}_{P_{a,x}}[\ell(h, u; Y, a, x)]$ for $P_{A,X}$-almost every (a, x).

Proposition 4 establishes that solving the pointwise dual problem in Eq. (42) is equivalent to finding the global minimizer $(h^\star, u^\star)$ of the risk function in Def. 4. This amortization substantially improves tractability: instead of solving a separate optimization for each (a, x), we learn functional parameters that generalize across the covariate space.

Since the risk function in Eq. (43) depends on the unknown propensity score e, we must estimate it from data. However, estimating e introduces errors that can propagate into the bound estimates. To mitigate this, we construct a debiased risk function that achieves first-order insensitivity (Neyman-orthogonality) to perturbations in e:

Definition 5 (Debiased Risk). Let $\eta_f'(A, X)$ be the first-order derivative of $\eta_f(A, X)$ w.r.t. e. The debiased risk function is

$$\mathcal{R}^{\mathrm{db}}(\beta, \gamma; e) \triangleq \mathbb{E}[\ell^{\mathrm{db}}(V; (\beta, \gamma), e)], \tag{46}$$

where

$$\ell^{\mathrm{db}}(V; (\beta, \gamma), e) \triangleq \underbrace{\exp(h_\beta(A, X))\Big\{\eta_f(A, X) + g^*\big(\tfrac{\varphi(Y) - u_\gamma(A,X)}{\exp(h_\beta(A,X))}\big)\Big\} + u_\gamma(A, X)}_{\text{Eq. (44)}} \tag{47}$$

$$+ \sum_{a \in \mathcal{A}} e_a(X) \exp(h_\beta(a, X))\eta_f'(a, X)\left(\mathbf{1}(A = a) - e_a(X)\right) \tag{48}$$

Here, Eq. (48) is an error correction terms, which makes $\ell^{\mathrm{db}}(V; (\beta, \gamma), e)$ invariant to the small perturbation to first-order perturbations in e (i.e., *Neyman-orthogonal* (Chernozhukov et al., 2018)):

Lemma 1 (Orthogonality). For any direction functions $\{s_a(\cdot)\}_{a\in\mathcal{A}}$ and any perturbation path $e_{t,a} \triangleq e_a + ts_a$ with sufficiently small $|t|$, $\frac{\partial}{\partial t}\mathcal{R}^{\text{db}}(\beta,\gamma;e_t)\big|_{t=0} = 0$ for all (β,γ).

We now present our estimation procedure based on cross-fitting:

Definition 6 (Debiased Causal Bound Estimators). Fix a functional φ and an f-divergence. Let ℓ^{db} and $\mathcal{R}^{\text{db}}$ be as in Def. 5. The debiased estimator of the upper causal bound $\theta_u(a,x)$ for any $(a,x)\in\mathcal{A}\times\mathcal{X}$ is constructed as follows:

1. Randomly split the dataset $\mathcal{D}$ (with size n) into K disjoint folds $\mathcal{D}_1,\cdots,\mathcal{D}_K$.
2. For each k fold, learn $\widehat{e}_a^k$ using $\mathcal{D}_{-k}\triangleq\mathcal{D}\setminus\mathcal{D}_k$ for all $a\in\mathcal{A}$.
3. For each fold k, solve $\widehat{\vartheta}_k \triangleq (\hat{\beta}_k,\hat{\gamma}_k)\in\arg\min_{\beta,\gamma}\sum_{i|V_i\in\mathcal{D}_k}\ell^{\text{db}}(V_i;(\beta,\gamma),\widehat{e}^k)$.
4. For each fold k, evaluate
$$\widehat{h}_k(a,x)\triangleq h_{\widehat{\beta}_k}(a,x), \qquad \widehat{u}_k(a,x)\triangleq u_{\widehat{\gamma}_k}(a,x), \tag{49}$$
$$\widehat{\lambda}_k(a,x)\triangleq\exp\{\widehat{h}_k(a,x)\}, \qquad \widehat{\eta}_f^k(a,x)\triangleq B_f(\widehat{e}_a^k(x)). \tag{50}$$
5. For each fold k and each $i\in\mathcal{D}_k$, evaluate $Z_i^k \triangleq g^*\big(\frac{\varphi(Y_i)-\widehat{u}_k(A_i,X_i)}{\widehat{\lambda}_k(A_i,X_i)}\big)$, and learn a regressor $\widehat{m}_k$ by regressing Z_i^k onto (A,X) using $\mathcal{D}_k$.
6. Evaluate $\widehat{\theta}_{\text{up}}^{(k)}(a,x)\triangleq\widehat{\lambda}_k(a,x)(\widehat{\eta}_f^k(a,x)+\widehat{m}_k(a,x))+\widehat{u}_k(a,x)$ and return $\widehat{\theta}_{\text{up}}(a,x)\triangleq(1/K)\sum_{k=1}^K\widehat{\theta}_{\text{up}}^{(k)}(a,x)$.

We now analyze the error of the proposed debiased estimator under following set of assumptions:

Assumption 2 (Regularity-1). Let $e\triangleq\{e_a(\cdot): a\in\mathcal{A}\}$ be the true propensity score, $\vartheta\triangleq(\beta,\gamma)$ and $\vartheta_0\triangleq(\beta_0,\gamma_0)\in\arg\min_{\beta,\gamma}\mathcal{R}^{\text{db}}(\beta,\gamma;e)$.

1. **Positivity:** $e_a(x)\in[c,1-c]$ for some constant $0<c<1/2$ for all $a,x\in\mathcal{A}\times\mathcal{X}$.
2. **f-divergence regularity:** f is convex and twice continuously differentiable; and the induced radius $B_f(e_a(x))$ is twice continuously differentiable on $[c,1-c]$, with bounded first and second derivative; i.e., $\sup_{e\in[c,1-c]}|B_f(e)|+|B_f'(e)|+|B_f''(e)|<\infty$.
3. **Loss regularity:** For each fixed $e\in[c,1-c]$, the map $\vartheta\mapsto\ell^{\text{db}}(V;\vartheta,e)$ is twice continuously differentiable, with
$$\sup_{\vartheta,e}\|\ell^{\text{db}}(V;\vartheta,e)\|_2^2<\infty,\quad \sup_{\vartheta,e}\|\nabla_\theta\ell^{\text{db}}(V;\vartheta,e)\|_2^2<\infty,\quad \sup_{\theta,e}\|\nabla^2_{\vartheta\vartheta}\ell^{\text{db}}(V;\vartheta,e)\|_2^2<\infty.$$
4. **Higher-order smoothness:** Let $H(\vartheta;e)\triangleq\nabla^2_{\vartheta\vartheta}R^{\text{db}}(\vartheta;e)$. There exists a neighborhood Θ_0 of ϑ containing ϑ_0 and constants $0<\kappa\le\kappa_2<\infty$ such that
$$\kappa_1\mathbf{I}\ \preceq\ H(\vartheta;e)\ \preceq\ \kappa_2\mathbf{I}\quad\text{for all }\vartheta\in\Theta_0. \tag{51}$$
5. **Uniform LLN:** For each fold k, define the empirical risk w.r.t. ℓ^{db} with the training

fold is $\widehat{R}_k^{\mathrm{db}}(\vartheta, \widehat{e}^k)$. Then, we have a uniform law-of-large-number:

$$\sup_{\vartheta} |\widehat{R}_k^{\mathrm{db}}(\vartheta; \widehat{e}^k) - R^{\mathrm{db}}(\vartheta; \widehat{e}^k)| = O_p(n^{-1/2}), \tag{52}$$

Assumption 3 (Regularity-2). For $\vartheta \triangleq (\beta, \gamma)$, let $Z_\vartheta \triangleq g^*(\frac{\varphi(Y)-u_\gamma(A,X)\}}{\exp(h_\beta(A,X))})$ and $m_\vartheta(a,x) \triangleq \mathbb{E}[Z_\vartheta \mid A = a, X = x]$. Let $\widehat{m}_k$ be the estimate for m_ϑ using the k-fold data.

1. **Bounded nuisances:** $h_{\beta_0}, u_{\gamma_0}, h_{\widehat{\beta}_k}, u_{\widehat{\gamma}_k}$ are bounded by some constant M.
2. **Lipschitz parameterization:** The map $(\beta, \gamma) \mapsto (h_\beta(a,x), u_\gamma(a,x))$ is Lipschitz in $\vartheta = (\beta, \gamma)$ uniformly over all (a,x) with constant L_ϑ; i.e.,
$$\sup_{a,x} (|h_\beta(a,x) - h_{\beta'}(a,x)| + |u_\gamma(a,x) - u_{\gamma'}(a,x)|) \leq L_\vartheta \|\vartheta - \vartheta'\|. \tag{53}$$
3. **Smoothness of g^*:** The convex conjugate g^* is continuously differentiable with bounded derivative; i.e., $\sup_{t\in\mathcal{T}} |(g^*)'(t)| < \infty$ where $\mathcal{T}$ is a range where $g^*(t)$ is well-defined.
4. **Assumption on regression:** $\|\widehat{m}_k - m_{\widehat{\vartheta}_k}\|_2 = O_p(s_n)$, where s_n is some sequence $s_n \to 0$; There exists a constant L_m s.t. $\|m_\vartheta - m_{\vartheta'}\|_2 \leq L_m \|\vartheta - \vartheta'\|$.
5. **Correct model choice**: Let $\overline{\theta}^\star_{\varphi,0}(a,x) \triangleq \mathbb{E}[\ell(V; (\beta_0, \gamma_0), e) \mid A = a, X = x]$. Then $\overline{\theta}^\star_{\varphi,0}(a,x) \triangleq \mathbb{E}[\ell(V; (\beta_0, \gamma_0), e) \mid A = a, X = x] = \overline{\theta}_\varphi(a,x)$ for all (a,x).

We now formalize the convergence rate of the proposed debiased estimator:

Theorem 3 (Error Analysis). Under Assumption 2, fix a fold k. Let e be the true propensity score, and $\vartheta_0 \triangleq (\beta_0, \gamma_0) \in \arg\min_\vartheta \mathcal{R}^{\mathrm{db}}(\vartheta; e)$ for $\vartheta \triangleq (\beta, \gamma)$. Let $\widehat{\vartheta}_k \triangleq (\widehat{\beta}_k, \widehat{\gamma}_k)$ be the minimizer from Step 3 in Def. 6 with $\widehat{e}^k$. Define $r_n \triangleq O_P(\|\widehat{e}^k - e\|_2)$. Then,

$$\|\widehat{\vartheta}_k - \vartheta_0\|_2^2 = O_p(n^{-1/2} + r_n^2). \tag{54}$$

Furthermore, let $Z_\vartheta \triangleq g^*(\frac{\varphi(Y)-u_\gamma(A,X)}{\exp(h_\beta(A,X))})$ and $m_\vartheta(a,x) \triangleq \mathbb{E}[Z_\vartheta \mid A = a, X = x]$. Define $s_n \triangleq O_P(\|\widehat{m}_k - m_{\widehat{\vartheta}_k}\|_2)$ where $\widehat{m}_k$ is from Step 5 in Def. 6. Let $\widehat{\theta}_{\mathrm{up}}^{(k)}$ be the estimated upper causal bound for the fold k. Under additional Assumption 3,

$$\|\widehat{\theta}_{\mathrm{up}}^{(k)} - \theta_{\mathrm{up}}\|_2^2 = O_p(n^{-1/2} + r_n^2 + s_n^2). \tag{55}$$

Thm. 3 demonstrates the sample efficiency of our debiased estimator. Even when the nuisance components (the propensity score and the pseudo-outcome regression) converge slowly (e.g., at rate $n^{-1/4}$), both the dual parameters $\widehat{\vartheta}_k$ and the upper-bound estimator $\widehat{\theta}_{\mathrm{up}}^{(k)}$ achieve the faster rate (e.g., at rate $n^{-1/2}$). Specifically, the sample-efficiency gain is of order $O_P(r_n^2)$ rather than $O_P(r_n)$ that would result from using the non-debiased risk function (Eq. (43)). This improvement stems from the orthogonal construction of the debiased risk, which eliminates first-order sensitivity to propensity score errors (Lemma 1). Consequently, nuisance components can be estimated using flexible machine learning methods while the estimator retains faster convergence rates.

5.1. Ensemble Bound Aggregation

Different f-divergences encode distinct notions of distributional discrepancy, and no single divergence uniformly dominates others in tightness across all data distributions. Consequently, we estimate bounds using a collection $\mathcal{F}$ of f-generators (e.g., $\mathcal{F} = \{f_{\mathrm{KL}}, f_{\mathrm{TV}}, f_{\chi^2}, \cdots\}$), yielding a family of upper bound estimates $\widehat{\boldsymbol{\theta}}_{\mathrm{up}} \triangleq \{\widehat{\theta}_{\mathrm{up},f} : f \in \mathcal{F}\}$ where $\widehat{\theta}_{\mathrm{up},f}$ is an upper-bound estimate for a fixed $f \in \mathcal{F}$. Let $\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}$ be defined similarly. To construct the tightest valid interval, we aggregate these bounds while accounting for potential finite-sample violations due to estimation error and numerical instability.

Our aggregation strategy addresses this challenge via order statistics:

Definition 7 (**k-th order statistics aggregator**). Let $\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}, \widehat{\boldsymbol{\theta}}_{\mathrm{up}}$ denote candidate lower and upper bounds, respectively, with $n_f \triangleq |\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}| = |\widehat{\boldsymbol{\theta}}_{\mathrm{up}}|$. For $k \in \{1, \dots, n_f\}$, the *k-th order-statistics aggregator* (k-agg) is defined as the pair $(\widehat{\theta}^k_{\mathrm{lo}}, \widehat{\theta}^k_{\mathrm{up}})$, where $\widehat{\theta}^k_{\mathrm{lo}}$ is the k-th largest element of $\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}$ and $\widehat{\theta}^k_{\mathrm{up}}$ is the k-th smallest element of $\widehat{\boldsymbol{\theta}}_{\mathrm{up}}$.

The following lemma formalizes the validity condition for the k-th order aggregator:

Lemma 2 (**Valid Coverage under Partial Correctness**). For a fixed (a, x),

- $\widehat{\theta}^k_{\mathrm{lo}}(a, x) \le \theta(a, x)$ iff at least $(n_f - k + 1)$ elements of $\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}$ are smaller or equal to $\theta(a, x)$.
- $\widehat{\theta}^k_{\mathrm{up}}(a, x) \ge \theta(a, x)$ iff at least $(n_f - k + 1)$ elements of $\widehat{\boldsymbol{\theta}}_{\mathrm{up}}$ are greater or equal to $\theta(a, x)$.

Lemma 2 guarantees that the k-agg produces valid bounds as long as at least $(n_f - k + 1)$ divergences yield correct estimates. This robustness property is critical: even if a minority of divergences fail (due to finite-sample violations or numerical issues), the aggregator automatically discards outliers by selecting the k-th order statistic. In practice, the k-agg is implemented by initializing $k = 1$ (selecting the tightest bounds) and iteratively incrementing $k \leftarrow k+1$ until $\widehat{\theta}^k_{\mathrm{lo}} \le \widehat{\theta}^k_{\mathrm{up}}$ is satisfied.

5.2. Debiased Estimation for Average Causal Effects

When covariates are absent ($X = \emptyset$), the estimation procedure simplifies substantially. The marginal propensity score $e_a \triangleq \Pr(A = a)$ can be estimated at rate $o_P(n^{-1/2})$ via sample proportions, eliminating the need for the debiasing correction in Eq. (48). We now specialize our framework to this covariate-free setting.

Definition 8 (**Risk Function (Marginal Case)**). Let $h \triangleq \{h_a \in \mathbb{R}^+ : a \in \mathcal{A}\}$ and $u \triangleq \{u_a \in \mathbb{R}^+ : a \in \mathcal{A}\}$. Let $V \triangleq (A, Y)$ and $\eta^a_f \triangleq B_f(e_a)$. A risk function for causal bound when $X = \emptyset$ is

$$\mathcal{R}(h, u; e) \triangleq \mathbb{E}_P[\ell(V; (h, u), \eta_f)], \tag{56}$$

where

$$\ell(V; (h, u), e) \triangleq \exp(h_A)\{\eta^a_f + g^\star(\tfrac{\varphi(Y) - u_A}{\exp(h_A)})\} + u_A. \tag{57}$$

The estimator for the marginal case directly minimizes the risk in Def. 8 without debiasing:

Definition 9 (**Bound Estimator (Marginal Case)**). Fix a functional φ and an f-divergence. Let ℓ and $\mathcal{R}$ be as in Def. 8. Let the observed sample be i.i.d. $\{V_i \triangleq (A_i, Y_i)\}_{i=1}^n$. Define $n_a \triangleq \sum_{i=1}^n \mathbf{1}(A_i = a)$. The estimator of the upper causal bound $\overline{\theta}_\varphi(a)$ for any $a \in \mathcal{A}$ is constructed as follows:

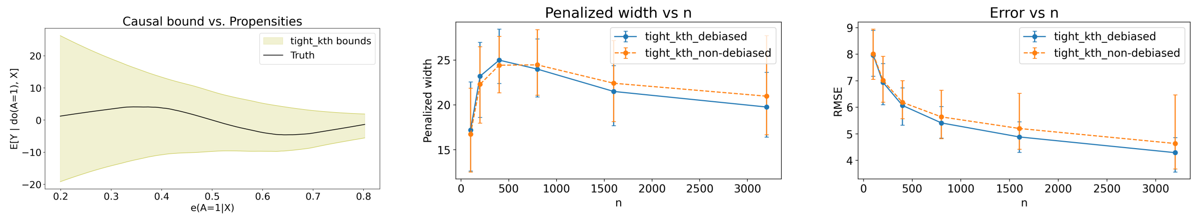
1. Estimate the marginal propensity $\widehat{e}_a \triangleq n_a/n$.
2. Solve $\widehat{\vartheta} \triangleq (\widehat{h}, \widehat{u}) \in \arg\min_{h,u} \sum_{i=1}^n \ell(V_i; (h, u), \widehat{e})$.
3. Evaluate $\widehat{\lambda}_a \triangleq \exp(\widehat{h}_a)$.
4. Define the pseudo-outcome $\widehat{Z}_i \equiv g^*\Big(\frac{\phi(Y_i) - \widehat{u}_{A_i}}{\widehat{\lambda}_{A_i}}\Big)$ and evaluate $\widehat{m}_a \triangleq (1/n_a) \sum_{i:A_i=a} \widehat{Z}_i$.
5. Return $\widehat{\theta}_{\varphi,f}(a) \equiv \widehat{\lambda}_a \Big(\widehat{\eta}_{f,a} + \widehat{m}_a\Big) + \widehat{u}_a$, for $a \in \mathcal{A}$.

We now analyze the error of the proposed debiased estimator under following set of assumptions:

Assumption 4 (**Regularity (Marginal Case)**). Let $e \triangleq \{e_a : a \in \mathcal{A}\}$ where $e_a \triangleq \Pr(A = a)$, $\vartheta \triangleq (\beta, \gamma)$ and $\vartheta_0 \in \arg\min_\vartheta \mathcal{R}(\vartheta; e)$ where $\vartheta \triangleq (h, u) \triangleq \{(h_a, u_a) : a \in \mathcal{A}\}$. Let $Z_\vartheta \triangleq g^*(\frac{\varphi(Y) - u_A}{\exp(h_A)})$. Let $m_{\vartheta,a} \triangleq \mathbb{E}_{P_a}[Z_\vartheta]$.

1. **Positivity:** $e_a \in [c, 1-c]$ for some constant $0 < c < 1/2$ for all $a \in \mathcal{A}$.
2. **f-divergence regularity:** f is convex and twice continuously differentiable; and the induced radius B_f is twice continuously differentiable on $[c, 1-c]$ with bounded derivatives; i.e., $\sup_{e\in[c,1-c]} |B_f(e)| + |B_f'(e)| + |B_f''(e)| < \infty$.
3. **Loss regularity:** For each fixed $e \in [c, 1-c]$, the map $\vartheta \mapsto \ell(V; \vartheta, e)$ is twice continuously differentiable, with
$$\sup_{\vartheta,e} \|\ell(V;\vartheta,e)\|_2^2 < \infty, \quad \sup_{\vartheta,e} \|\nabla_\theta \ell(V;\vartheta,e)\|_2^2 < \infty, \quad \sup_{\theta,e} \|\nabla^2_{\vartheta\vartheta} \ell(V;\vartheta,e)\|_2^2 < \infty.$$
4. **Higher-order smoothness:** Let $H(\vartheta; e) \triangleq \nabla^2_{\vartheta\vartheta} R(\vartheta; e)$. There exists a neighborhood Θ_0 of ϑ containing ϑ_0 and constants $0 < \kappa \le \kappa_2 < \infty$ such that
$$\kappa_1 \mathbf{I} \ \preceq\ H(\vartheta; e) \ \preceq\ \kappa_2 \mathbf{I} \quad \text{for all } \vartheta \in \Theta_0. \tag{58}$$
5. **Uniform LLN:** Define the empirical risk w.r.t. ℓ with the training fold is $\widehat{R}(\vartheta, \widehat{e})$. Then, we have a uniform law-of-large-number:
$$\sup_\vartheta |\widehat{R}(\vartheta; \widehat{e}) - R(\vartheta; \widehat{e})| = O_p(n^{-1/2}). \tag{59}$$
6. **Bounded parameters:** h_a, u_a are bounded by some constant M.
7. **Smoothness of g^*:** The convex conjugate g^* is continuously differentiable with bounded derivative; i.e., $\sup_{t\in\mathcal{T}} |(g^*)'(t)| < \infty$ where $\mathcal{T}$ is a range where $g^*(t)$ is well-defined.

The following theorem establishes the convergence rate for the marginal case:



(a) (b) (c)

Figure 2: **(a)** Bounds vs. propensity scores; **(b)** Penalized width vs. sample size, where the penalized width is p-width $\triangleq$ width $\times (1 + a \times \max(0, (1-\alpha) - \text{coverage}))$ with $a = 10$ and $\alpha = 0.95$; **(c)** Convergence rate comparison

Theorem 4 (Error Analysis (Marginal Case)). Assume Assumption 4. Let $e_0 \triangleq \{e_{0,a} : a \in \mathcal{A}\}$ with $e_{0,a} \equiv \Pr(A = a)$ and let $\widehat{e}_a \equiv n_a/n$. Let $\vartheta_0 \in \arg\min_{\vartheta \in \Theta} R(\vartheta; e_0)$ and $\widehat{\vartheta} \in \arg\min_{\vartheta \in \Theta} \widehat{R}_n(\vartheta; \widehat{e})$. Let $\overline{\theta}_\varphi$ and $\widehat{\theta}_\varphi$ be the population target and the estimator defined in Def. 9 (marginal case). Then

$$\|\widehat{\vartheta} - \vartheta_0\|_2^2 = O_p(n^{-1/2}), \qquad \|\widehat{\theta}_\varphi - \overline{\theta}_\varphi\|_2^2 = O_p(n^{-1/2}). \tag{60}$$

Thm. 4 shows that in the marginal case, both the dual parameters and the bound estimator achieve a *squared* error rate of $O_p(n^{-1/2})$ (implying a parameter convergence rate of $O_p(n^{-1/4})$) without requiring debiasing. This is because the marginal propensity score $\widehat{e}_a = n_a/n$ converges at rate $O_P(n^{-1/2})$, which is fast enough that first-order bias terms vanish asymptotically. This contrasts with the conditional case (Thm. 3), where debiasing is essential to handle slower convergence rates of nonparametric nuisance estimators.

6. Experiments

This section empirically validates our framework across both synthetic and real-world data. Our goal is to bound the conditional causal mean $\theta(1, x) \triangleq \mathbb{E}[Y \mid \mathrm{do}(A = 1), X = x]$ using our proposed bounds in Def. 6.

Across all experiments, we estimate the propensity score via XGBoost (Chen and Guestrin, 2016) and fit the dual functions $\lambda(a, x) = \exp(h(a, x))$ and $u(\cdot)$ using a neural network trained with two-fold cross-fitting. We consider the f-divergences in Cor. T1.1 (KL, Jensen–Shannon, Hellinger, TV, and χ^2), and the order-statistics aggregator (Def. 7). In the figures below, the label `tight_kth` denotes the aggregated interval with $k = 5$.

Synthetic data experiments. We generate synthetic data from the SCM in Fig. 1a with $X \in \mathbb{R}^5$, binary treatment $A \in \{0, 1\}$, and a continuous outcome Y with heavy-tail noise following a Student's t-distribution with 3 degrees of freedom, which has substantially thicker tails than the standard normal distribution. Fig. 2a demonstrates the validity of our method: the true effect curve for $\theta(1, x)$ lies within the estimated `tight_kth` bounds across all propensity score regimes, even under heavy-tailed noise. This plot also shows how interval width shrinks as $e_a(x) \to 1$, as expected from our theory. Notably, the χ^2 divergence consistently produces the tightest bounds among all f-divergences considered.

We next examine the debiasing benefit formalized in Thm. 3. Fig. 2b compares penalized width, defined as p-width $\triangleq$ width $\times (1 + a \times \max(0, (1 - \alpha) - \text{coverage}))$ with $a = 10$ and

$\alpha = 0.95$, where coverage denotes the fraction of evaluation points $\{x_1, \cdots, x_n\}$ satisfying $\theta(1, x_i) \in [\widehat{\theta}_{\mathrm{lo}}(1, x_i), \widehat{\theta}_{\mathrm{up}}(1, x_i)]$, between debiased and non-debiased estimators. As expected, the debiased estimator achieves tighter penalized width as n increases, reflecting improved finite-sample efficiency. Fig. 2c further illustrates robustness to nuisance estimation error: we add convergence noise $\epsilon \sim \mathcal{N}(n^{-1/4}, n^{-1/4})$ to the estimated propensity score and compare convergence rates using an oracle estimator equipped with the true propensity score. The debiased estimator maintains its convergence rate despite slower propensity score estimation, confirming the theoretical guarantee of first-order insensitivity.

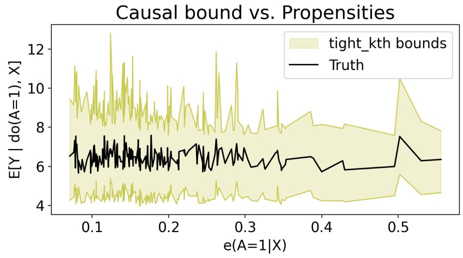


Figure 3: IHDP data analysis

Semi-synthetic IHDP benchmark. We also validate our method on the well-known IHDP (Infant Health and Development Program) benchmark (Hill, 2011; Louizos et al., 2017; AMLab Amsterdam, 2020). This dataset originates from a randomized trial studying the effect of home visits by specialists on future cognitive test scores, with confounders $X \in \mathbb{R}^{25}$ capturing characteristics of the children and their mothers. Following Louizos et al. (2017), we de-randomize the treatment assignment to introduce confounding. In our experiment, we observe only five covariates and treat the remaining 20 as hidden confounders. Fig. 3 confirms that our bounds tightly contain the true causal effect $\mathbb{E}[Y \mid \mathrm{do}(A = 1), X = x]$ across the full range of estimated propensity scores $\widehat{e}_1(x)$ for $x \in \mathcal{X}$.

7. Conclusion

This paper develops an information-theoretic framework for partial identification of causal effects under unmeasured confounding. The key contribution is deriving data-driven bounds on f-divergences between observational and interventional distributions using only the propensity score, without requiring auxiliary variables or user-specified sensitivity parameters. These divergence bounds translate into causal effect bounds that simultaneously address four key limitations of existing methods: (1) accommodating unbounded continuous outcomes, (2) avoiding full structural causal model specification, (3) providing heterogeneous effect bounds conditional on covariates, and (4) achieving computational tractability through debiased semiparametric estimation. Our debiased semiparametric estimators achieve $\sqrt{n}$-consistency even when nuisance components converge at slower nonparametric rates, leveraging Neyman-orthogonality to eliminate first-order bias. Experiments on synthetic and semi-synthetic benchmarks confirm valid coverage across propensity score regimes and demonstrate robustness to heavy-tailed outcome distributions. Future work includes extending the framework to continuous treatments and deriving sharper bounds by incorporating additional structural information or auxiliary data.

References

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

AMLab Amsterdam (2020). Amlab-amsterdam/cevae: Causal effect inference with deep latent-variable models. GitHub repository. Archived July 17, 2020.

Balazadeh Meresht, V., Syrgkanis, V., and Krishnan, R. G. (2022). Partial identification of treatment effects with implicit generative models. *Advances in Neural Information Processing Systems*, 35:22816–22829.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pages 46–54. Elsevier.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Dorn, J. and Guo, K. (2023). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657.

Ghassami, A., Shpitser, I., and Tchetgen, E. T. (2023). Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hu, Y., Wu, Y., Zhang, L., and Wu, X. (2021). A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12104–12112.

Jin, Y., Ren, Z., and Zhou, Z. (2022). Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*.

Kitagawa, T. (2021). The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*, 225(2):231–253. Themed Issue: Treatment Effect 1.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, pages 1071–1102.

Levis, A. W., Bonvini, M., Zeng, Z., Keele, L., and Kennedy, E. H. (2025). Covariate-assisted bounds on causal effects with instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N., and Shalit, U. (2023). B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, pages 26599–26618. PMLR.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. (2023). Stochastic causal programming for bounding treatment effects. In *Conference on Causal Learning and Reasoning*, pages 142–176. PMLR.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York. 2nd edition, 2009.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Sachs, M. C., Jonzon, G., Sjölander, A., and Gabriel, E. E. (2023). A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576.

Semenova, V. (2025). Generalized lee bounds. *Journal of Econometrics*, 251:106055.

Shridharan, M. and Iyengar, G. (2023). Scalable computation of causal bounds. *Journal of Machine Learning Research*, 24(237):1–35.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947.

Tan, J., Blanchet, J., and Syrgkanis, V. (2024). Consistency of neural causal partial identification. *Advances in Neural Information Processing Systems*, 37:68956–68999.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313.

Xia, K. M., Pan, Y., and Bareinboim, E. (2022). Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations.*

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of statistics*, 50(5):2587.

Zhang, J. and Bareinboim, E. (2021). Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215.

Supplement of Data-Driven Information-Theoretic Causal Bounds under Unmeasured Confounding

A. Simulation Details

This section provides the technical specifications for the synthetic and semi-synthetic experiments presented in Section 6.

A.1. Synthetic Data Generating Process

We consider a nonlinear structural causal model (SCM) where the observational data (X, A, Y) is generated as follows. The features $X \in \mathbb{R}^d$ are sampled from a uniform distribution: $X_j \sim \text{Uniform}(-2, 2)$ for $j = 1, \dots, d$. A latent confounder U is sampled from a standard normal distribution, $U \sim \mathcal{N}(0, 1)$.

The treatment assignment A follows a Bernoulli distribution with a propensity score dependent on both X and U. Specifically, we define the logits L as:

$$L(X, U) = X^\top w + 0.8U + 0.5 \sin(X_0) - 0.25X_0^2, \tag{61}$$

where $w \in \mathbb{R}^d$ is a fixed weight vector sampled from $\mathcal{N}(0, 0.6^2)$. The treatment is assigned as $A \sim \text{Bernoulli}(e(X, U))$, with

$$e(X, U) = 0.05 + 0.9 \cdot \text{sigmoid}(L(X, U)), \tag{62}$$

which ensures overlap by constraining the propensity score to $[0.05, 0.95]$.

The outcome Y is generated as:

$$Y = \mu(X) + \tau(X, e(X, U))A + 0.7U + \epsilon, \tag{63}$$

where ϵ follows a Student's t-distribution with 3 degrees of freedom ($\epsilon \sim t_3(0, 1)$) to introduce heavy-tailed noise. The functions $\mu(X)$ and $\tau(X, p)$ are defined to capture complex nonlinearities and heterogeneity:

$$\mu(X) = 0.5 + 0.8 \tanh(X_0) + 0.25X_1^2 - 0.15 \sin(X_2), \tag{64}$$

$$\tau(X, p) = 0.7 + 0.2 \sin(X_0) + 0.1X_0 + 0.8(p - 0.5), \tag{65}$$

where p is the treatment assignment probability. This SCM introduces both selection bias via U and heterogeneous treatment effects that depend on the propensity score.

A.2. Neural Network Architecture and Training

For estimating the dual functions and the nuisance components (outcome models and propensity scores), we use Multi-Layer Perceptrons (MLPs) and XGBoost.

Architecture for Dual Functions The dual functions h are parameterized by an MLP with two hidden layers of 64 units each, using ReLU activations. We apply a clipping operation to the output of the dual network such that $h(X) \in [-20, 20]$ to ensure numerical stability during optimization. No dropout is used.

Optimization and Hyperparameters The dual networks are trained using the Adam optimizer with a learning rate of 5×10^{-4} and weight decay of 1×10^{-4}. We employ 2-fold cross-fitting to avoid overfitting and ensure the validity of the debiased estimator. For each fold, we train the dual network for 256 epochs. Early stopping with a patience of 10 epochs (monitored on a 20% validation split of the training fold) is used to prevent overtraining.

Nuisance Models Propensity scores and outcome means are estimated using XGBoost with the following hyperparameters:

- **Number of estimators**: 300 for propensity, 400 for outcome.
- **Maximum depth**: 10.
- **Learning rate**: 0.005.
- **Subsample / Colsample**: 0.8.

A.3. IHDP Benchmark Details

The IHDP benchmark is a semi-synthetic dataset based on a real-world randomized trial from the Infant Health and Development Program. We use the version where selection bias is introduced by removing a non-random subset of the treated group.

In our experiments, we treat 5 of the 25 covariates as observed and the remaining 20 as hidden confounders to simulate a scenario with unmeasured confounding. The evaluation is performed on a fixed set of units to compare the estimated bounds against the ground truth interventional effects provided by the benchmark. Training is conducted for 200 epochs for the IHDP-specific experiments.

B. Proofs

Proof of Thm. 1

We first declare some useful results:

Lemma 3 (f-divergence with Conditional Measure). Let P on $(\Omega, \mathcal{F})$ be an arbitrary probability measure. Let $E \in \mathcal{F}$ be a fixed event such that $P(E) = p \in (0, 1)$. Let $P_E(\cdot) \triangleq P(\cdot \mid E)$. Then,

$$D_f(P_E \| P) = pf(\tfrac{1}{p}) + (1-p)f(0) \tag{66}$$

Proof. Define the conditional-on-an-event measure P_E by

$$P_E(B) \coloneqq P(B \mid E) = \frac{P(B \cap E)}{P(E)}, \qquad \forall B \in \mathcal{F}, \tag{67}$$

where $P(E) = p \in (0,1)$. Then $P_E \ll P$ since $P(B) = 0 \Rightarrow P(B \cap E) = 0 \Rightarrow P_E(B) = 0$. Hence, by the Radon–Nikodým theorem, there exists a measurable function $g = \frac{dP_E}{dP}$ (unique P-a.e.) such that

$$P_E(B) = \int_B g(\omega)\, P(d\omega), \qquad \forall B \in \mathcal{F}. \tag{68}$$

A valid version is $g(\omega) = \frac{1}{p}\, \mathbf{1}_E(\omega)$ since, for any $B \in \mathcal{F}$,

$$\int_B \frac{1}{p} \mathbf{1}_E(\omega)\, P(d\omega) = \frac{1}{p} P(B \cap E) = \frac{P(B \cap E)}{P(E)} = P_E(B). \tag{69}$$

Therefore,

$$D_f(P_E \| P) = \int_\Omega f\left(\frac{dP_E}{dP}(\omega)\right) P(d\omega) \tag{70}$$

$$= \int_E f\left(\frac{1}{p}\right) P(d\omega) + \int_{E^c} f(0)\, P(d\omega) \tag{71}$$

$$= pf\left(\frac{1}{p}\right) + (1-p) f(0). \tag{72}$$

□

Lemma 4 (**Data Processing Inequality** (Csiszár, 1967))**.** Let P_X and Q_X denote probability measures on $(\mathcal{X}, \mathcal{F}_X)$. Let $P_{Y|X}$ be a Markov kernel from $(\mathcal{X}, \mathcal{F}_X)$ to $(\mathcal{Y}, \mathcal{F}_Y)$. Let P_Y, Q_Y be the transformation of P_X, Q_X, respectively, when pushed through $P_{Y|X}$; i.e., $P_Y(B) = \int_{\mathcal{X}} P_{Y|X}(B \mid x) dP_X(x)$, and Q_Y is defined similarly. Then, for any f-divergence, we have

$$D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X). \tag{73}$$

For any fixed $X = x$, define the event $E \coloneqq \{A = a\}$ under the measure $P_{U,A|X=x}$, so that $P(E \mid x) = P(A = a \mid X = x) = e_a(x)$. Let

$$P_E(\cdot \mid x) \coloneqq P_{U,A|X=x}(\,\cdot \mid E) = P_{U,A|X=x,A=a}. \tag{74}$$

By Lemma 3,

$$D_f\left(P_{U,A|X=x,A=a} \,\|\, P_{U,A|X=x}\right) = e_a(x) f\left(\frac{1}{e_a(x)}\right) + (1 - e_a(x)) f(0) \equiv B_f(e_a(x)). \tag{75}$$

Define the (Markov) transition kernel $K_{a,x}$ from $(\mathcal{U} \times \mathcal{A}, \mathcal{F}_{U,A})$ to $(\mathcal{Y}, \mathcal{F}_Y)$ by, for any $B \in \mathcal{F}_Y$,

$$K_{a,x}(B \mid u, a') \coloneqq P(Y \in B \mid U = u, A = a, X = x), \tag{76}$$

(note $K_{a,x}$ is constant in a').

Pushing $P_{U,A|X=x}$ through $K_{a,x}$ yields

$$\int K_{a,x}(B \mid u, a')\, P_{U,A|X=x}(du\, da') = \int P(Y \in B \mid u, a, x)\, P_{U,A|X=x}(du\, da') = P(Y \in B \mid \mathrm{do}(A = a), X = x). \tag{77}$$

Similarly, pushing $P_{U,A|X=x,A=a}$ through $K_{a,x}$ yields

$$\int K_{a,x}(B \mid u, a')\, P_{U,A|X=x,A=a}(du\, da') = \int P(Y \in B \mid u, a, x)\, P_{U|X=x,A=a}(du) = P(Y \in B \mid A = a, X = x). \tag{78}$$

By the data processing inequality (Lemma 4),

$$D_f\Big(P_{Y|A=a,X=x} \,\|\, P_{Y|\mathrm{do}(A=a),X=x}\Big) \le D_f\Big(P_{U,A|X=x,A=a} \,\|\, P_{U,A|X=x}\Big) = B_f(e_a(x)). \tag{79}$$

■.

Proof of Cor. T1.1

KL. With $f(t) = t \log t$ with $f(0) = 0$, we have

$$B(e_a(x), f) = -e_a(x)\frac{1}{e_a(x)} \log e_a(x) = -\log e_a(x). \tag{80}$$

Therefore,

$$D_{\mathrm{KL}}(P(Y \mid a, x) \| Q(Y \mid a, x)) \le -\log e_a(x). \tag{81}$$

Hellinger. With $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$ with $f(0) = 1/2$, we have

$$\begin{aligned} B(e_a(x), f) &= e_a(x) f(\tfrac{1}{e_a(x)}) + (1 - e_a(x)) f(0) && (82)\\ &= \frac{1}{2} e_a(x)\Big(\sqrt{\tfrac{1}{e_a(x)}} - 1\Big)^2 + \frac{1}{2}(1 - e_a(x)) && (83)\\ &= 1 - \sqrt{e_a(x)}. && (84) \end{aligned}$$

To tighten, we use the following lemma:

Lemma 5 (Hellinger divergence vs. KL divergence). For any P, Q such that $P \ll Q$,

$$D_{\mathrm{H}}(P \| Q) \le \frac{1}{2} D_{\mathrm{KL}}(P \| Q). \tag{85}$$

Proof. We start with

$$D_{\mathrm{H}}(P \| Q) \triangleq \frac{1}{2} \int (\sqrt{p}(x) - \sqrt{q}(x))^2 dx = 1 - \int \sqrt{p(x) q(x)} dx. \tag{86}$$

Define $\mathtt{BC}(P, Q) \triangleq \int \sqrt{p(x) q(x)} dx$. Then, $D_{\mathrm{H}}(P \| Q) = 1 - \mathtt{BC}(P, Q)$. Define $D_{\mathrm{B}}(P \| Q) \triangleq -\log \mathtt{BC}(P, Q)$, which is known as Bhattacharyya distance.

Define $r(X) \triangleq \frac{q(x)}{p(x)}$. Then,

$$\mathtt{BC}(P, Q) = \int \sqrt{p(x)q(x)}dx = \int \sqrt{\tfrac{q(x)}{p(x)}}p(x)dx = \mathbb{E}_P\Big[\sqrt{r(X)}\Big]. \tag{87}$$

By Jensen's inequality, we have

$$\log \mathtt{BC}(P, Q) = \log \mathbb{E}_P\Big[\sqrt{r(X)}\Big] \geq \mathbb{E}_P\Big[\log \sqrt{r(X)}\Big] = \frac{1}{2}\mathbb{E}_P[\log r(X)]. \tag{88}$$

Also,

$$\mathbb{E}_P[\log r(X)] = \int p(x) \log \frac{q(x)}{p(x)} dx = -D_{\mathrm{KL}}(P, Q). \tag{89}$$

Combining,

$$-\frac{1}{2}D_{\mathrm{KL}}(P\|Q) \leq \log \mathtt{BC}(P, Q) \quad \Leftrightarrow \quad 1 - \exp\Big(-\frac{1}{2}D_{\mathrm{KL}}(P\|Q)\Big) \geq 1 - \mathtt{BC}(P, Q). \tag{90}$$

Finally,

$$D_{\mathrm{H}}(P\|Q) = 1 - \mathtt{BC}(P, Q) \leq 1 - \exp\Big(-\frac{1}{2}D_{\mathrm{KL}}(P\|Q)\Big) \leq \frac{1}{2}D_{\mathrm{KL}}(P\|Q), \tag{91}$$

where the last inequality holds since $1 - e^{-u} \leq u$ for any $u \geq 0$. □

As a result, we can derive

$$D_{\mathrm{H}}(P(Y \mid a, x)\|Q(Y \mid a, x)) \leq -\frac{1}{2} \log e_a(x). \tag{92}$$

Finally, for $e_a(x) \in (0, 1)$, the following holds:

$$1 - \sqrt{e_a(x)} \leq -\frac{1}{2} \log e_a(x). \tag{93}$$

χ^2-divergence. Set $f(t) \triangleq \frac{1}{2}(t-1)^2$. Then, $B_f(e_a) = \frac{1-e_a(x)}{2e_a(x)}$.

Total variation. First, $B_{f_{\mathrm{TV}}}(e) = 1 - e$.

Second, by Pinsker's inequality and the above inequality,

$$D_{\mathrm{TV}}(P\|Q) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)} \leq \sqrt{-\frac{1}{2}\log e_a(x)}. \tag{94}$$

By Bretagnolle–Huber bound (Bretagnolle and Huber, 1979) and the above inequality,

$$D_{\mathrm{TV}}(P\|Q) \leq \sqrt{1 - \exp(-D_{\mathrm{KL}}(P\|Q))} \leq \sqrt{1 - e_a(x)}. \tag{95}$$

Finally, $\min\Big(1 - e_a(x), \sqrt{1 - e_a(x)}, \sqrt{-\frac{1}{2}\log e_a(x)}\Big) = 1 - e_a(x)$ for all $e_a(x) \in (0, 1)$.

Jensen-Shannon. With $f_{\mathrm{JS}}(t) \triangleq \frac{1}{2}(t\log t - (t+1)\log(\frac{t+1}{2}))$ and $f_{\mathrm{JS}}(0) = \frac{1}{2}\log 2$, we have:

$$B_{f_{\mathrm{JS}}}(e_a(x)) = e_a(x) f_{\mathrm{JS}}\Big(\frac{1}{e_a(x)}\Big) + (1 - e_a(x)) f_{\mathrm{JS}}(0) \tag{96}$$

$$= \frac{e_a(x)}{2}\left[\frac{1}{e_a(x)}\log\Big(\frac{1}{e_a(x)}\Big) - \Big(\frac{1}{e_a(x)} + 1\Big)\log\left(\frac{1+e_a(x)}{2e_a(x)}\right)\right] + \frac{1-e_a(x)}{2}\log 2 \tag{97}$$

$$= \frac{1}{2}\left[-\log e_a(x) - (1+e_a(x))\log\left(\frac{1+e_a(x)}{2e_a(x)}\right) + (1-e_a(x))\log 2\right] \tag{98}$$

$$= \frac{1}{2}\left[-\log e_a(x) - (1+e_a(x))\left[\log(1+e_a(x)) - \log e_a(x) - \log 2\right] + \log 2 - e_a(x)\log 2\right] \tag{99}$$

$$= \frac{1}{2}\Big[-\log e_a(x) - (1+e_a(x))\log(1+e_a(x)) + \log e_a(x) + e_a(x)\log e_a(x) + 2\log 2 + e_a(x)\log 2 - e_a(x)\log 2\Big] \tag{100}$$

$$= \frac{1}{2}\left[e_a(x)\log e_a(x) - (1+e_a(x))\log(1+e_a(x)) + 2\log 2\right] \tag{101}$$

$$= \frac{1}{2}\log\left(\frac{4e_a(x)^{e_a(x)}}{(1+e_a(x))^{1+e_a(x)}}\right). \tag{102}$$

■

Proof of Cor. T1.3

By definition, for any class of functions $\mathcal{F}$, the Integral Probability Metric (IPM) satisfies:

$$D_{\mathrm{IPM},\mathcal{F}}(P\|Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_P[f(Y)] - \mathbb{E}_Q[f(Y)]\right|. \tag{103}$$

If $f(Y) \in [a, b]$ for all $y \in \mathcal{Y}$, then for any probability measures P, Q:

$$|\mathbb{E}_P[f(Y)] - \mathbb{E}_Q[f(Y)]| \le (b-a) D_{\mathrm{TV}}(P, Q). \tag{104}$$

For $\mathcal{F}_C \triangleq \{f : \|f\|_\infty < C\}$, we have $f(y) \in (-C, C)$, so the range is $2C$. Consequently,

$$D_{\mathrm{IPM},\mathcal{F}_C}(P_{a,x}\|Q_{a,x}) \le 2C \cdot D_{\mathrm{TV}}(P_{a,x}\|Q_{a,x}). \tag{105}$$

From Corollary T1.1, we have $D_{\mathrm{TV}}(P_{a,x}\|Q_{a,x}) \le 1 - e_a(x)$. Furthermore, by Pinsker's inequality and the KL bound from Corollary T1.1:

$$D_{\mathrm{TV}}(P_{a,x}\|Q_{a,x}) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_{a,x}\|Q_{a,x})} \le \sqrt{-\frac{1}{2}\log e_a(x)}. \tag{106}$$

Combining these yields the result for IPM.

For MMD, let $\mathcal{H}_k$ be an RKHS with kernel $\mathtt{k}$ such that $\mathtt{k}(y, y) \le K$ for all y. For any $h \in \mathcal{H}_k$ with $\|h\|_{\mathcal{H}_k} \le 1$, we have $|h(y)| = |\langle h, \mathtt{k}_y\rangle| \le \|h\|_{\mathcal{H}_k}\sqrt{\mathtt{k}(y, y)} \le \sqrt{K}$. Thus, $h(y) \in [-\sqrt{K}, \sqrt{K}]$, and the range is $2\sqrt{K}$. Following similar logic:

$$D_{\mathrm{MMD},\mathtt{k}}(P_{a,x}\|Q_{a,x}) \le 2\sqrt{K} \cdot D_{\mathrm{TV}}(P_{a,x}\|Q_{a,x}). \tag{107}$$

Using the TV bounds derived above, we obtain the MMD bound. ■

Proof of Prop. 1

We will prove the following statement: For any arbitrary function f over some space $\mathcal{X}$, the following holds: $\inf_{x\in\mathcal{X}} f(x) \;=\; -\sup_{x\in\mathcal{X}} \big(-f(x)\big)$.

$$\inf_{x\in\mathcal{X}} f(x) \;=\; -\sup_{x\in\mathcal{X}} \big(-f(x)\big). \tag{108}$$

For any $x \in \mathcal{X}$,

$$-f(x) \le -\inf_{x'} f(x'), \;\; \forall x \in \mathcal{X} \implies \inf_{x\in\mathcal{X}} f(x) \le -\sup_{x\in\mathcal{X}}(-f(x)). \tag{109}$$

Also, by the definition of infimum, for any $\varepsilon > 0$, there exists x_ϵ such that

$$f(x_\varepsilon) \le \inf_{x\in\mathcal{X}} f(x) + \varepsilon. \tag{110}$$

Then,

$$-f(x_\varepsilon) \ge -\inf_{x} f(x) - \varepsilon \implies \sup_{x\in\mathcal{X}}(-f(x)) \ge -\inf_{x\in\mathcal{X}} f(x) - \varepsilon. \tag{111}$$

By taking $\epsilon \downarrow 0$, we have $\sup_{x\in\mathcal{X}}(-f(x)) \ge -\inf_{x\in\mathcal{X}} f(x)$. The proof is done by combining these two inequalities. ■

Proof of Thm. 2

Fix (a, x) and write $P_{a,x}$ and $Q_{a,x}$ for the observational and interventional laws on $(\mathcal{Y}, \mathcal{F})$. By Assumption 1 (mutual absolute continuity), the Radon–Nikodym derivative

$$s(y) \;:=\; \frac{dQ_{a,x}}{dP_{a,x}}(y) \tag{112}$$

exists and satisfies $s(Y) > 0$ $P_{a,x}$-a.s. For any measurable ϕ with $\mathbb{E}_{Q_{a,x}}[|\phi(Y)|] < \infty$,

$$\mathbb{E}_{Q_{a,x}}[\phi(Y)] = \int \phi(y)\, Q_{a,x}(dy) = \int \phi(y)\, s(y)\, P_{a,x}(dy) = \mathbb{E}_{P_{a,x}}[s(Y)\phi(Y)]. \tag{113}$$

Moreover, $\mathbb{E}_{P_{a,x}}[s(Y)] = \int dQ_{a,x} = 1$. Define $g(s) := sf(1/s)$ for $s > 0$. Then

$$\mathbb{E}_{P_{a,x}}[g(s(Y))] = \int s(y) f(1/s(y))\, P_{a,x}(dy) = \int f\left(\frac{dP_{a,x}}{dQ_{a,x}}(y)\right) Q_{a,x}(dy) = D_f(P_{a,x}\|Q_{a,x}). \tag{114}$$

Hence the constraint $D_f(P_{a,x}\|Q_{a,x}) \le \eta_f(a, x)$ is equivalent to $\mathbb{E}_{P_{a,x}}[g(s(Y))] \le \eta_f(a, x)$, and the upper bound admits the primal form

$$\theta_{\mathrm{up}}(a, x) = \sup_{s>0} \Big\{ \mathbb{E}_{P_{a,x}}[s(Y)\phi(Y)] : \mathbb{E}_{P_{a,x}}[s(Y)] = 1,\; \mathbb{E}_{P_{a,x}}[g(s(Y))] \le \eta_f(a, x) \Big\}. \tag{115}$$

This is a convex optimization problem (equivalently, minimize $-\mathbb{E}_{P_{a,x}}[s\phi]$) with an affine equality and a convex inequality constraint. Slater's condition holds because $s(\cdot) \equiv 1$ is feasible and satisfies $\mathbb{E}_{P_{a,x}}[g(1)] = f(1) = 0 < \eta_f(a,x)$ (for $\eta_f(a,x) > 0$). Therefore, strong duality applies and the optimal value equals the dual optimal value.

Introduce Lagrange multipliers $u \in \mathbb{R}$ for $\mathbb{E}_{P_{a,x}}[s] = 1$ and $\lambda \geq 0$ for $\mathbb{E}_{P_{a,x}}[g(s)] \leq \eta_f(a,x)$. The Lagrangian is

$$\mathcal{L}(s,\lambda,u) = \mathbb{E}_{P_{a,x}}[s(Y)\phi(Y)] + u(1 - \mathbb{E}_{P_{a,x}}[s(Y)]) + \lambda(\eta_f(a,x) - \mathbb{E}_{P_{a,x}}[g(s(Y))]), \tag{116}$$

i.e.

$$\mathcal{L}(s,\lambda,u) = u + \lambda\eta_f(a,x) + \mathbb{E}_{P_{a,x}}\big[s(Y)(\phi(Y) - u) - \lambda g(s(Y))\big]. \tag{117}$$

Thus

$$\theta_{\mathrm{up}}(a,x) = \inf_{\lambda\geq 0,\ u\in\mathbb{R}} \sup_{s>0} \mathcal{L}(s,\lambda,u). \tag{118}$$

For $\lambda > 0$, define $t(Y) \coloneqq (\phi(Y) - u)/\lambda$. Using separability of the integrand in $s(\cdot)$ and the standard interchange theorem for integral functionals (equivalently, the conjugate-of-integral identity), we have

$$\sup_{s>0} \mathbb{E}_{P_{a,x}}\big[s(Y)t(Y) - g(s(Y))\big] = \mathbb{E}_{P_{a,x}}\Big[\sup_{s>0}\{s\,t(Y) - g(s)\}\Big] = \mathbb{E}_{P_{a,x}}\big[g^*(t(Y))\big], \tag{119}$$

where $g^*(t) \coloneqq \sup_{s>0}\{st - g(s)\}$ is the convex conjugate of g. Consequently,

$$\sup_{s>0} \mathcal{L}(s,\lambda,u) = u + \lambda\eta_f(a,x) + \lambda\,\mathbb{E}_{P_{a,x}}\left[g^*\left(\frac{\phi(Y)-u}{\lambda}\right)\right]. \tag{120}$$

Minimizing over (λ,u) yields the stated dual representation:

$$\theta_{\mathrm{up}}(a,x) = \inf_{\lambda>0,\ u\in\mathbb{R}} \left\{\lambda\eta_f(a,x) + u + \lambda\,\mathbb{E}_{P_{a,x}}\left[g^*\left(\frac{\phi(Y)-u}{\lambda}\right)\right]\right\}. \tag{121}$$

■

Proof of Prop. 2

Substitute $r = 1/s$. Then, $st - g(s) = st - sf(1/s) = \frac{t-f(r)}{r}$. Taking $\sup_{s>0}$ is the same as taking $\sup_{r>0}$. Therefore, $g^*(t) = \sup_{r>0} \frac{t-f(r)}{r}$.

For the optimality condition, define

$$H_t(r) \coloneqq \frac{t - f(r)}{r}, \qquad r > 0. \tag{122}$$

Assume the supremum is attained at some $r^* > 0$, and set

$$v \coloneqq g^*(t) = H_t(r^*) = \frac{t - f(r^*)}{r^*}. \tag{123}$$

Then for every $r > 0$,

$$\frac{t - f(r)}{r} \leq v \quad\Longleftrightarrow\quad f(r) \geq t - vr. \tag{124}$$

At $r = r^*$ we have equality: $f(r^*) = t - vr^*$. Hence for all $r > 0$,

$$f(r) \geq f(r^*) - v(r - r^*) = f(r^*) + a(r - r^*), \tag{125}$$

where $a := -v$. By the supporting-hyperplane characterization of the convex subdifferential, this implies $a \in \partial f(r^*)$. Finally, $f(r^*) = t - vr^*$ gives

$$t = f(r^*) - r^* a, \qquad g^*(t) = v = -a. \tag{126}$$

If f is differentiable at r^*, then $\partial f(r^*) = \{f'(r^*)\}$ and the conclusion follows.

■

Proof of Coro. P1

KL. $f_{\mathrm{KL}}(r) = r \log r$. Then,

$$g_{\mathrm{KL}}(s) = s f_{\mathrm{KL}}(1/s) = s \cdot \frac{1}{s} \log(1/s) = -\log s. \tag{127}$$

Now, compute $g^*_{\mathrm{KL}}(t) = \sup_{s>0}\{st + \log s\}$. Let $\psi(s) = st + \log s$. Then, $\psi'(s) = t + 1/s$. If $t < 0$, the stationary point is $s^* = -1/t > 0$, giving $g^*_{\mathrm{KL}}(t) = \psi(s^*) = (-1/t)t + \log(-1/t) = -1 - \log(-t)$. If $t \geq 0$, then $st + \log s \to \infty$ as $s \to \infty$, so $g^*_{\mathrm{KL}}(t) = +\infty$.

Hellinger. $f_H(r) = \frac{1}{2}(\sqrt{r} - 1)^2 = \frac{1}{2}(r - 2\sqrt{r} + 1)$. Then,

$$g_H(s) = s f_H(1/s) = \frac{1}{2}\, s\Big(\frac{1}{s} - \frac{2}{\sqrt{s}} + 1\Big) = \frac{1}{2}(1 - 2\sqrt{s} + s). \tag{128}$$

Note $g^*_H(t) = \sup_{s>0}\Big\{st - \frac{1}{2}(1 - 2\sqrt{s} + s)\Big\}$. Let $u \triangleq \sqrt{s} > 0$ so $s = u^2$. The objective becomes $F(u) = tu^2 - \frac{1}{2}(1 - 2u + u^2) = \Big(t - \frac{1}{2}\Big)u^2 + u - \frac{1}{2}$. If $t < 1/2$, F is concave quadratic in u. Since $F'(u) = 2(t - 1/2)u + 1$, $u^* = \frac{1}{1-2t}$. Plugging in,

$$g^*_H(t) = F(u^*) = \Big(t - \tfrac{1}{2}\Big)\frac{1}{(1-2t)^2} + \frac{1}{1-2t} - \frac{1}{2} = \frac{t}{1-2t}. \tag{129}$$

If $t \geq 1/2$, then $F(u) \to \infty$ as $u \to \infty$, so $g^*_{\mathrm{H}}(t) = +\infty$.

χ^2**.** $g_{\chi^2}(s) = s f_{\chi^2}(1/s) = \frac{1}{2}\, s\Big(\frac{1}{s} - 1\Big)^2 = \frac{(1-s)^2}{2s}$. Also, $g^*_{\chi^2}(t) = \sup_{s>0}\Big\{st - \frac{(1-s)^2}{2s}\Big\}$, where $\frac{(1-s)^2}{2s} = \frac{1}{2}\Big(\frac{1}{s} - 2 + s\Big)$. Then, the objective is

$$st - \frac{1}{2}\Big(\frac{1}{s} - 2 + s\Big) = 1 + s\Big(t - \tfrac{1}{2}\Big) - \frac{1}{2s}. \tag{130}$$

Differentiate w.r.t. s:

$$\frac{d}{ds}\Big(1 + s(t - \tfrac{1}{2}) - \frac{1}{2s}\Big) = (t - \tfrac{1}{2}) + \frac{1}{2s^2}. \tag{131}$$

Plugging in (using $1/s^* = \sqrt{1-2t}$):

$$g^*_{\chi^2}(t) = 1 + s^*(t - \tfrac{1}{2}) - \frac{1}{2s^*} = 1 - \frac{\sqrt{1-2t}}{2} - \frac{\sqrt{1-2t}}{2} = 1 - \sqrt{1-2t}. \tag{132}$$

At $t = 1/2$, this becomes 1. If $t > 1/2$, the term $s(t - \frac{1}{2})$ drives the supremum to $+\infty$ as $s \to \infty$.

TV. $g_{\mathrm{TV}}(s) = sf_{\mathrm{TV}}(1/s) = \frac{1}{2}\, s\left|\frac{1}{s} - 1\right| = \frac{1}{2}|1-s|$. Also, $g^*_{\mathrm{TV}}(t) = \sup_{s>0}\left\{st - \frac{1}{2}|1-s|\right\}$. Split this into two regions, where $s \geq 1$ and $0 < s \leq 1$.

When $s \geq 1$, $|1-s| = s-1$. So,

$$st - \tfrac{1}{2}(s-1) = s\Big(t - \tfrac{1}{2}\Big) + \tfrac{1}{2}. \tag{133}$$

If $t > 1/2$: this goes to $+\infty$ as $s \to \infty$. If $t \leq 1/2$, the maximum over $s \geq 1$ occurs at the smallest s; i.e., $s = 1$, giving value t.

When $0 < s \leq 1$, $|1-s| = 1-s$, so

$$st - \tfrac{1}{2}(1-s) = s\Big(t + \tfrac{1}{2}\Big) - \tfrac{1}{2}. \tag{134}$$

If $t < -1/2$, then its maximum is $-1/2$. If $t \geq -1/2$, then it's maximized at $s = 1$, giving value t. As a result,

$$g^*_{\mathrm{TV}}(t) = \begin{cases} -\frac{1}{2}, & \text{if } t \leq -\frac{1}{2}, \\ t, & \text{if } -\frac{1}{2} < t \leq \frac{1}{2}, \\ +\infty, & \text{if } t > \frac{1}{2}. \end{cases} \tag{135}$$

■

Jensen-Shannon. $g_{\mathrm{JS}}(s) = \frac{1}{2}\Big(s\log s - (1+s)\log(1+s) + (1+s)\log 2\Big)$. To compute $g^*_{\mathrm{JS}}(t) = \sup_{s>0}\{st - g_{\mathrm{JS}}(s)\}$, let $F(s) = st - g_{\mathrm{JS}}(s)$.

We have

$$g'_{\mathrm{JS}}(s) = \frac{1}{2}\Big(\log s - \log(1+s) + \log 2\Big) = \frac{1}{2}\log\Big(\frac{2s}{1+s}\Big). \tag{136}$$

Set $F'(s) = 0$, which means $t = g'_{\mathrm{JS}}(s)$; i.e.,

$$2t = \log\Big(\frac{2s}{1+s}\Big) \quad \Longleftrightarrow \quad e^{2t} = \frac{2s}{1+s}. \tag{137}$$

Solving this for s gives

$$e^{2t}(1+s) = 2s \;\Rightarrow\; s^* = \frac{e^{2t}}{2 - e^{2t}}. \tag{138}$$

This requires $2 - e^{2t} > 0$, i.e., $t < \frac{1}{2}\log 2$. If $t \geq \frac{1}{2}\log 2$, the objective grows like $s(t - \frac{1}{2}\log 2)$ for large s, hence the supremum is $+\infty$.

Now evaluate the objective at s^*. Let $z \triangleq e^{2t}$ so that $s^* = z/(2-z)$ and $1 + s^* = 2/(2-z)$.

Then,

$$\log s^* = \log z - \log(2-z), \qquad \log(1+s^*) = \log 2 - \log(2-z). \tag{139}$$

Plug into $g_{\mathrm{JS}}(s)$:

$$g_{\mathrm{JS}}(s^*) = \frac{1}{2}\Big(s^* \log s^* - (1+s^*)\log(1+s^*) + (1+s^*)\log 2\Big) = \frac{1}{2}\Big(s^* \log z + \log(2-z)\Big). \tag{140}$$

Since $\log z = 2t$, this is $g_{\mathrm{JS}}(s^*) = ts^* + \frac{1}{2}\log(2-e^{2t})$. Therefore,

$$g^*_{\mathrm{JS}}(t) = s^* t - g_{\mathrm{JS}}(s^*) = -\frac{1}{2}\log(2-e^{2t}), \qquad \text{for } t < \tfrac{1}{2}\log 2, \tag{141}$$

and $g^*_{\mathrm{JS}}(t) = +\infty$ otherwise. ■

Proof of Prop. 4

((1) $\implies$ (2)). For each fixed (a,x), define

$$\Delta(a,x) \triangleq \ell(h^\star, u^\star; a, x) - \operatorname*{ess\,inf}_{h,u\in\mathcal{F}} \ell(h,u;a,x). \tag{142}$$

Assume, for contradiction, that (2) fails; i.e., $P_{A,X}(B) > 0$ for $B \triangleq \{(a,x) : \Delta(a,x) > 0\}$. By the definition of the essential infimum and the decomposability of $\mathcal{F}$, there exists a measurable pair $(\tilde{h}, \tilde{u}) \in \mathcal{F}$ such that $\ell(\tilde{h}, \tilde{u}; a, x) < \ell(h^\star, u^\star; a, x)$ on a set of positive measure $B' \subseteq B$.

Define $h'(a,x) \triangleq \tilde{h}(a,x)\mathbf{1}((a,x) \in B') + h^\star(a,x)\mathbf{1}((a,x) \notin B')$ and define $u'(a,x)$ similarly. By the decomposability assumption, $(h', u') \in \mathcal{F}$. Then,

$$\begin{aligned}
\mathcal{R}(h^\star, u^\star) &= \mathbb{E}[\ell(h^\star, u^\star, A, X)\mathbf{1}((A,X) \notin B')] + \mathbb{E}[\ell(h^\star, u^\star, A, X)\mathbf{1}((A,X) \in B')] && (143)\\
&> \mathbb{E}[\ell(h^\star, u^\star, A, X)\mathbf{1}((A,X) \notin B')] + \mathbb{E}[\ell(\tilde{h}, \tilde{u}, A, X)\mathbf{1}((A,X) \in B')] && (144)\\
&= \mathcal{R}(h', u'). && (145)
\end{aligned}$$

This contradicts the optimality of $(h^\star, u^\star)$ in (1). Therefore, $P_{A,X}(B) = 0$; i.e., $(h^\star, u^\star)$ is a minimizer of $\ell(h,u;a,x)$ for $P_{A,X}$-almost every (a,x).

((2) $\implies$ (1)). Since $\ell(h^\star, u^\star; a, x) \le \ell(h,u;a,x)$ for all $(h,u) \in \mathcal{F}$ and for $P_{A,X}$-almost every (a,x), integrating yields $\mathcal{R}(h^\star, u^\star) \le \mathcal{R}(h,u)$ for all $(h,u) \in \mathcal{F}$. ■

Proof of Lemma 1

Define

$$\begin{aligned}
e_a &\triangleq \Pr(A = a \mid X) && (146)\\
\lambda_a &\triangleq \exp(h_\beta(a, X)) && (147)\\
B_a(e) &\triangleq B_f(e_a(X)) && (148)\\
u_a &\triangleq u(a, X) && (149)\\
g^*_a &\triangleq g^*\left(\frac{\varphi(Y) - u(A,X)}{\lambda_A(X)}\right). && (150)
\end{aligned}$$

Then,

$$\ell(V;(\beta,\gamma),e) \triangleq \lambda_A(B_A + g_A^*) + u_A(X). \tag{151}$$

Define

$$L_1(e) \triangleq \lambda_A(B_A + g_A^*) + u_A(X). \tag{152}$$

The correction term is

$$L_2(e) \triangleq \sum_a e_a \lambda_a B_a' \{\mathbf{1}(A=a) - e_a\}. \tag{153}$$

Then, $\mathcal{R}^{\text{db}}(e) \triangleq \mathbb{E}[L_1(e) + L_2(e)]$. Then,

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}[L_1(e_t)]\Big|_{t=0} &= \frac{\partial}{\partial t}\mathbb{E}[L_1(e_A + ts_A)]\Big|_{t=0} && (154)\\ &= \mathbb{E}\left[\lambda_A B'(e_A)s_A\right]. && (155)\end{aligned}$$

Also,

$$L_2(e) \triangleq \sum_a \underbrace{e_a \lambda_a B_a'}_{U_a(e_a)} \underbrace{\{\mathbf{1}(A=a) - e_a\}}_{V_a(e_a)}. \tag{156}$$

Then,

$$\frac{\partial}{\partial t}\mathbb{E}[L_2(e_t)]\Big|_{t=0} = \mathbb{E}\left[\sum_{a\in\mathcal{A}}\left(\frac{\partial U_a}{\partial t}V_a + \frac{\partial V_a}{\partial t}U_a\right)\right], \tag{157}$$

where

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}} U_a'(e)V_a(e_a)\right] = \mathbb{E}_X\left[\sum_{a\in\mathcal{A}} U_a'(e)\mathbb{E}_{A|X}[\mathbf{1}(A=a) - e_a]\right] = 0, \tag{158}$$

and

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}} U_a(e)V_a'(e_a)\right] = -\mathbb{E}\left[\sum_{a\in\mathcal{A}} U_a(e)s_a\right] = -\mathbb{E}_X\left[\sum_{a\in\mathcal{A}} e_a\lambda_a B_a' s_a\right]. \tag{159}$$

Then,

$$\begin{aligned}\frac{\partial R^{\text{db}}}{\partial t} &= \mathbb{E}[\lambda_A B'(e_A)s_A] - \mathbb{E}_X\left[\sum_{a\in\mathcal{A}} e_a\lambda_a B_a' s_a\right] && (160)\\ &= \mathbb{E}_X\left[\sum_{a\in\mathcal{A}} e_a\lambda_a B_a' s_a\right] - \mathbb{E}_X\left[\sum_{a\in\mathcal{A}} e_a\lambda_a B_a' s_a\right] && (161)\\ &= 0. && (162)\end{aligned}$$

■

Proof of Theorem 3

Lemma 6 (Higher-order smoothness ⇒ Local quadratic expansion inequality). Higher-order smoothness in Assumption 2 implies the local quadratic expansion inequality:

$$\frac{\kappa_1}{2}\|\vartheta-\vartheta_0\|^2 \;\le\; \mathcal{R}^{\mathrm{db}}(\vartheta;e_0)-\mathcal{R}^{\mathrm{db}}(\vartheta_0;e_0) \;\le\; \frac{\kappa_2}{2}\|\vartheta-\vartheta_0\|^2,\;\text{for } \vartheta\in\Theta_0. \tag{163}$$

Proof of Lemma 6. Let $r(t)\triangleq\mathcal{R}(\vartheta_t;e_0)$, where $\vartheta_t\triangleq\vartheta_0+t(\vartheta-\vartheta_0)$ for $t\in[0,1]$. By Taylor's theorem with integral remainder,

$$r(1)=r(0)+r'(0)+\int_0^1(1-t)r''(t)dt. \tag{164}$$

Since ϑ_0 is a local minimizer, $r'(0)=(\vartheta-\vartheta_0)^\intercal\nabla_\vartheta\mathcal{R}(\vartheta_0;e_0)=0$. The second derivative is

$$r''(t)=(\vartheta-\vartheta_0)^\intercal H(\vartheta_t;e_0)(\vartheta-\vartheta_0). \tag{165}$$

Under the Higher-order smoothness assumption ($\kappa_1 I\preceq H(\vartheta;e_0)\preceq\kappa_2 I$ for $\vartheta\in\Theta_0$), and assuming convexity of Θ_0 so that the path lies in Θ_0, we have

$$\frac{\kappa_1}{2}\|\vartheta-\vartheta_0\|_2^2\le\int_0^1(1-t)(\vartheta-\vartheta_0)^\intercal H(\vartheta_t;e_0)(\vartheta-\vartheta_0)dt\le\frac{\kappa_2}{2}\|\vartheta-\vartheta_0\|_2^2. \tag{166}$$

□

Proof of Eq. (54)

For brevity, we write $R(\vartheta;e')\triangleq R^{\mathrm{db}}(\vartheta;e')$ for any ϑ and e'. Let $\widehat{R}_k$ denote the empirical risk of R using the k'th fold dataset.

We decompose the population excess risk using a telescoping sum:

$$R(\widehat{\vartheta}_k;e_0)-R(\vartheta_0;e_0)=\underbrace{R(\widehat{\vartheta}_k;e_0)-R(\widehat{\vartheta}_k;\widehat{e}^k)}_{(A)}+\underbrace{R(\widehat{\vartheta}_k;\widehat{e}^k)-\widehat{R}_k(\widehat{\vartheta}_k;\widehat{e}^k)}_{(B)} \tag{167}$$

$$+\underbrace{\widehat{R}_k(\widehat{\vartheta}_k;\widehat{e}^k)-\widehat{R}_k(\vartheta_0;\widehat{e}^k)}_{\le 0}+\underbrace{\widehat{R}_k(\vartheta_0;\widehat{e}^k)-R(\vartheta_0;\widehat{e}^k)}_{(C)} \tag{168}$$

$$+\underbrace{R(\vartheta_0;\widehat{e}^k)-R(\vartheta_0;e_0)}_{(D)}. \tag{169}$$

The term ≤ 0 is due to the optimality of $\widehat{\vartheta}_k$ for the empirical risk objective. We will show that

1. $(B)+(C)=O_p(n^{-1/2})$ by the uniform LLN in Assumption 2.
2. $(A)+(D)=O_p(r_n^2)$ by the orthogonality and smoothness in Assumption 2.

As a result,

$$R(\widehat{\vartheta}_k;e_0)-R(\vartheta_0;e_0)=O_p(n^{-1/2})+O_p(r_n^2). \tag{170}$$

Bounds for $(B)+(C)$. Terms $(B)+(C)$ are bounded by Uniform LLN as follows:

$$(B)+(C) \le 2\sup_{\vartheta} |R(\vartheta;\hat{e}^k) - \widehat{R}_k(\vartheta;\hat{e}^k)| = O_p(n^{-1/2}). \tag{171}$$

Bounds for $(A)+(D)$. Assume that the risk functional $e \mapsto R(\vartheta;e)$ is twice Fréchet differentiable with bounded second derivatives on the positivity region. Fix a ϑ. Consider a parametric submodel $t \mapsto e^t \triangleq e_0 + t(\hat{e}^k - e_0)$. Let $\delta e_0 \triangleq \hat{e}^k - e_0$.

By Taylor's theorem, there exists $e^\dagger$ between e_0 and $\hat{e}^k$ such that:

$$R(\vartheta;\hat{e}^k) = R(\vartheta;e_0) + \nabla_e R(\vartheta;e_0)[\delta e_0] + \frac{1}{2}\nabla_{ee} R(\vartheta;e^\dagger)[\delta e_0, \delta e_0]. \tag{172}$$

Rearranging for term (D) where $\vartheta = \vartheta_0$:

$$(D) = R(\vartheta_0;\hat{e}^k) - R(\vartheta_0;e_0) = \nabla_e R(\vartheta_0;e_0)[\delta e_0] + \frac{1}{2}\nabla_{ee} R(\vartheta_0;e^\dagger)[\delta e_0, \delta e_0]. \tag{173}$$

By Lemma 1 (Orthogonality), $\nabla_e R(\vartheta_0;e_0)[\delta e_0] = 0$. Using the boundedness of $\nabla_{ee} R$ (Assumption 2), we have $(D) = O_P(\|\hat{e}^k - e_0\|_2^2) = O_P(r_n^2)$.

For term (A) where $\vartheta = \widehat{\vartheta}_k$:

$$(A) = R(\widehat{\vartheta}_k;e_0) - R(\widehat{\vartheta}_k;\hat{e}^k) = -\nabla_e R(\widehat{\vartheta}_k;e_0)[\delta e_0] + O_P(r_n^2). \tag{174}$$

Crucially, Lemma 1 states that orthogonality holds for *all* ϑ (not just ϑ_0). Therefore, $\nabla_e R(\widehat{\vartheta}_k;e_0)[\delta e_0] = 0$ directly. This implies that the first-order error term vanishes exactly, and we are left only with the second-order remainder:

$$(A) = O_P(r_n^2). \tag{175}$$

Combining yields:

$$(A)+(D) = O_P(r_n^2). \tag{176}$$

Bound Derivation. Combining all terms:

$$R(\widehat{\vartheta}_k;e_0) - R(\vartheta_0;e_0) = O_p(n^{-1/2}) + O_P(r_n^2). \tag{177}$$

Assuming consistency (so $\widehat{\vartheta}_k \in \Theta_0$ w.h.p), we apply Lemma 6:

$$\frac{\kappa_1}{2}\|\widehat{\vartheta}_k - \vartheta_0\|^2 \le R(\widehat{\vartheta}_k;e_0) - R(\vartheta_0;e_0). \tag{178}$$

Solving the quadratic inequality for $\|\widehat{\vartheta}_k - \vartheta_0\|$ establishes:

$$\|\widehat{\vartheta}_k - \vartheta_0\|^2 = O_p(n^{-1/2} + r_n^2). \tag{179}$$

Proof of Eq. (55)

For brevity, we just write

$$\theta_k \triangleq \widehat{\theta}_\varphi^{(k)}, \quad \lambda_k \triangleq \widehat{\lambda}_k, \quad \eta_k \triangleq \widehat{\eta}_f^k, \quad m_k \triangleq \widehat{m}_k, \quad u_k \triangleq \widehat{u}_k. \tag{180}$$

All the true parameters are indexed as 0. For each (a, x),

$$\theta_k(a,x) - \overline{\theta}_\varphi(a,x) = \underbrace{(\lambda_k - \lambda_0)(\eta_0 + m_0)}_{(I)} + \underbrace{(\lambda_k - \lambda_0)(\eta_k - \eta_0)}_{(II)} \tag{181}$$

$$+ \underbrace{\lambda_0(\eta_k - \eta_0)}_{(III)} + \underbrace{\lambda_k(m_k - m_0)}_{(IV)} + \underbrace{(u_k - u_0)}_{(V)}. \tag{182}$$

We bound the squared L_2 norm of each term. By Lipschitz parametrization (Assumption 3):

$$\|(V)\|_2^2 = O_P(\|\widehat{\vartheta}_k - \vartheta_0\|_2^2). \tag{183}$$

For (I), using the boundedness of nuiances (Assumption 3, $|\eta_0 + m_0| \leq C$):

$$\|(I)\|_2^2 \leq C^2\|\lambda_k - \lambda_0\|_2^2 \leq C'\|\widehat{\vartheta}_k - \vartheta_0\|_2^2 = O_P(\|\widehat{\vartheta}_k - \vartheta_0\|_2^2). \tag{184}$$

For (III), using $|\lambda_0| \leq e^M$ and Lipschitz continuity of η (via B_f) with respect to e:

$$\|(\mathrm{III})\|_2^2 \leq e^{2M}\|\eta_k - \eta_0\|_2^2 = O_P(r_n^2). \tag{185}$$

For (II), we use the supremum bound on λ: $\|\lambda_k - \lambda_0\|_\infty \leq 2e^M$. Then:

$$\|(\mathrm{II})\|_2^2 \leq \|\lambda_k - \lambda_0\|_\infty^2 \|\eta_k - \eta_0\|_2^2 \leq 4e^{2M} r_n^2 = O_P(r_n^2). \tag{186}$$

Finally, consider (IV). Define $m_{\widehat{\varphi}}(a,x) \triangleq \mathbb{E}[Z_i^k \mid A = a, X = x]$. Decompose $m_k - m_0 = (m_k - m_{\widehat{\vartheta}}) + (m_{\widehat{\vartheta}} - m_0)$. By Assumption 3, $\|m_k - m_{\widehat{\vartheta}}\|_2 = O_P(s_n)$. By Lipschitz, $\|m_{\widehat{\vartheta}} - m_0\|_2 \leq L_m\|\widehat{\vartheta} - \vartheta_0\|$. Therefore,

$$\|(IV)\|_2^2 \leq e^{2M}(\|m_k - m_{\widehat{\vartheta}}\|_2 + \|m_{\widehat{\vartheta}} - m_0\|_2)^2 = O_P(s_n^2) + O_P(\|\widehat{\vartheta}_k - \vartheta_0\|^2). \tag{187}$$

Combining all terms shows that

$$\|\theta_k - \overline{\theta}_\varphi\|_2^2 = O_P(n^{-1/2} + r_n^2 + s_n^2). \tag{188}$$

Proof of Lemma 2

Let the sorted elements of $\widehat{\boldsymbol{\theta}}_{\mathrm{up}}$ be denoted by $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n_f)}$. By Definition 7, $\widehat{\theta}_{\mathrm{up}}^k = u_{(k)}$. The inequality $u_{(k)} \geq \theta$ holds if and only if at least $n_f - k + 1$ elements satisfy $u_i \geq \theta$ (since this is equivalent to having at most $k - 1$ elements strictly less than θ).

Similarly, let the sorted elements of $\widehat{\boldsymbol{\theta}}_{\mathrm{lo}}$ be $l_{(1)} \leq \cdots \leq l_{(n_f)}$. By definition, $\widehat{\theta}_{\mathrm{lo}}^k$ is the k-th largest element, which corresponds to $l_{(n_f-k+1)}$. The inequality $l_{(n_f-k+1)} \leq \theta$ holds if and only if at least $n_f - k + 1$ elements satisfy $l_i \leq \theta$ (since this is equivalent to having at most $k - 1$ elements strictly greater than θ). ■

Proof of Thm. 4

Write $\widehat{R} \equiv \widehat{R}_n$. Decompose the population excess risk:

$$\begin{aligned}0 \le R(\widehat{\vartheta}; e_0) - R(\vartheta_0; e_0) &= \underbrace{R(\widehat{\vartheta}; e_0) - R(\widehat{\vartheta}; \widehat{e})}_{(A)} + \underbrace{R(\widehat{\vartheta}; \widehat{e}) - \widehat{R}(\widehat{\vartheta}; \widehat{e})}_{(B)} \\ &\quad + \underbrace{\widehat{R}(\widehat{\vartheta}; \widehat{e}) - \widehat{R}(\vartheta_0; \widehat{e})}_{\le 0} + \underbrace{\widehat{R}(\vartheta_0; \widehat{e}) - R(\vartheta_0; \widehat{e})}_{(C)} + \underbrace{R(\vartheta_0; \widehat{e}) - R(\vartheta_0; e_0)}_{(D)}.\end{aligned}$$

By the uniform LLN in Assumption 4,

$$(B) + (C) \le 2 \sup_{\vartheta \in \Theta} \left|\widehat{R}(\vartheta; \widehat{e}) - R(\vartheta; \widehat{e})\right| = O_p(n^{-1/2}).$$

By Lipschitz continuity of $R(\vartheta; \cdot)$ in e uniformly over $\vartheta \in \Theta$,

$$|(A)| + |(D)| \le 2L_R \|\widehat{e} - e_0\|_1 = O_p(n^{-1/2}),$$

since $\widehat{e}_a = n_a/n$ implies $\|\widehat{e} - e_0\|_1 = O_p(n^{-1/2})$ under positivity.

Hence

$$0 \le R(\widehat{\vartheta}; e_0) - R(\vartheta_0; e_0) = O_p(n^{-1/2}).$$

Let Θ_0 be the neighborhood from the quadratic growth condition (Lemma 6). Since $R(\vartheta; e_0) - R(\vartheta_0; e_0)$ is bounded away from 0 on $\Theta \setminus \Theta_0$, the above display implies $\Pr(\widehat{\vartheta} \in \Theta_0) \to 1$. Therefore, on this event,

$$\frac{\kappa_1}{2}\|\widehat{\vartheta} - \vartheta_0\|_2^2 \le R(\widehat{\vartheta}; e_0) - R(\vartheta_0; e_0) = O_p(n^{-1/2}),$$

so $\|\widehat{\vartheta} - \vartheta_0\|_2^2 = O_p(n^{-1/2})$.

Next, write (as in Def. 9, marginal case)

$$\widehat{\theta}_\varphi(a) = \widehat{\lambda}_a(\widehat{\eta}_a + \widehat{m}_a) + \widehat{u}_a, \qquad \overline{\theta}_\varphi(a) = \lambda_{0,a}(\eta_{0,a} + m_{0,a}) + u_{0,a},$$

where $\lambda_a = \exp(h_a)$, $\eta_a = B_f(e_a)$, $Z_\vartheta \equiv g^*((\varphi(Y) - u_A)/\lambda_A)$, $m_{\vartheta,a} = \mathbb{E}[Z_\vartheta \mid A = a]$, and $\widehat{m}_a = n_a^{-1} \sum_{i:A_i=a} Z_{\widehat{\vartheta},i}$. Decompose, for each a,

$$\begin{aligned}\widehat{\theta}_\varphi(a) - \overline{\theta}_\varphi(a) &= (\widehat{\lambda}_a - \lambda_{0,a})(\eta_{0,a} + m_{0,a}) + (\widehat{\lambda}_a - \lambda_{0,a})(\widehat{\eta}_a - \eta_{0,a}) + \lambda_{0,a}(\widehat{\eta}_a - \eta_{0,a}) \\ &\quad + \widehat{\lambda}_a(\widehat{m}_a - m_{0,a}) + (\widehat{u}_a - u_{0,a}) \quad =: (I) + (II) + (III) + (IV) + (V).\end{aligned}$$

By boundedness of h and smoothness of $\exp(\cdot)$ on bounded sets, $\|\widehat{\lambda} - \lambda_0\|_2 \lesssim \|\widehat{h} - h_0\|_2 \le \|\widehat{\vartheta} - \vartheta_0\|_2$, and $\|\widehat{u} - u_0\|_2 \le \|\widehat{\vartheta} - \vartheta_0\|_2$. Thus $\|(I)\|_2^2 + \|(V)\|_2^2 = O_p(\|\widehat{\vartheta} - \vartheta_0\|_2^2) = O_p(n^{-1/2})$.

Also, $\|\widehat{\eta} - \eta_0\|_2 \lesssim \|\widehat{e} - e_0\|_1 = O_p(n^{-1/2})$ (bounded B_f'), so $\|(III)\|_2^2 = O_p(n^{-1})$. Moreover, $\|(II)\|_2 \le \|\widehat{\lambda} - \lambda_0\|_2 \|\widehat{\eta} - \eta_0\|_\infty = O_p(n^{-1/4}) \cdot O_p(n^{-1/2}) = O_p(n^{-3/4})$, hence $\|(II)\|_2^2 = O_p(n^{-3/2})$.

For (IV), decompose

$$\widehat{m}_a - m_{0,a} = \underbrace{\frac{1}{n_a} \sum_{i:A_i=a} (Z_{\widehat{\vartheta},i} - Z_{\vartheta_0,i})}_{(a)} + \underbrace{\left\{\frac{1}{n_a} \sum_{i:A_i=a} Z_{\vartheta_0,i} - \mathbb{E}[Z_{\vartheta_0} \mid A = a]\right\}}_{(b)} + \underbrace{(m_{\vartheta_0,a} - m_{\widehat{\vartheta},a})}_{(c)}.$$

By bounded derivative of g^* and bounded parameters, Z_ϑ is Lipschitz in ϑ, so $(a) = O_p(\|\widehat{\vartheta} - \vartheta_0\|_2) = O_p(n^{-1/4})$ and $(c) = O_p(\|\widehat{\vartheta} - \vartheta_0\|_2) = O_p(n^{-1/4})$. By positivity $n_a \asymp n$ and CLT, $(b) = O_p(n^{-1/2})$. Hence $\|\widehat{m} - m_0\|_2 = O_p(n^{-1/4})$. Since $\widehat{\lambda}$ is bounded, $\|(IV)\|_2^2 = O_p(n^{-1/2})$.

Collecting terms, the dominant squared contributions are $O_p(n^{-1/2})$ from (I), (IV), and (V), so $\|\widehat{\theta}_\varphi - \overline{\theta}_\varphi\|_2^2 = O_p(n^{-1/2})$. ■