Supplementary Material: Alleviating Privacy Attacks via Causal Learning

A Generalization Properties of Causal Models

A.1 Generalization over Different Distributions

We provide proofs for generalization properties of causal model over different distributions and a single datapoint.

Theorem 1. *Consider a structural causal graph* $\mathtt{G}$ *that connects* $\mathtt{X}$ *to* $\mathtt{Y}$*, and causal features* $\mathtt{X_C} \subseteq \mathtt{X}$ *where* $\mathtt{X_C}$ *represent the parents of* $\mathtt{Y}$ *under* $\mathtt{G}$*. Let* $\mathtt{P(X,Y)}$ *and* $\mathtt{P}^*\mathtt{(X,Y)}$ *be two distributions with arbitrary* $\mathtt{P(X)}$ *and* $\mathtt{P}^*\mathtt{(X)}$*, having overlap such that* $\mathtt{P(X=x)} > 0$ *whenever* $\mathtt{P}^*\mathtt{(X=x)} > 0$*. In addition, the causal relationship between* $\mathtt{X_C}$ *and* $\mathtt{Y}$ *is preserved, which implies that* $\mathtt{P(Y|X_C)} = \mathtt{P}^*\mathtt{(Y|X_C)}$*. Let* L *be a symmetric loss function that obeys the triangle inequality (such as L2 loss), and let* $\mathtt{f} : \mathtt{X_C} \rightarrow \mathtt{Y}$ *be the optimal predictor among all hypotheses using* $\mathtt{X_C}$ *features under* L*, i.e.,* $\mathtt{f} = \arg\min_{\mathtt{h}} \mathtt{L}_{\mathtt{x_c}}(\mathtt{y}, \mathtt{h}(\mathtt{x_c}))$ *for all* $\mathtt{x_c}$*, and thus* f *depends only on* $\Pr(\mathtt{Y|X_C})$ *(e.g.,* $\mathtt{f} := \mathbb{E}[\mathtt{Y|X_C}]$ *for L2 loss). Further, assume that* $\mathcal{H}_\mathtt{C}$ *represents the set of causal models* $\mathtt{h_c} : \mathtt{X_C} \rightarrow \mathtt{Y}$ *that may use all causal features and* $\mathcal{H}_\mathtt{A}$ *represent the set of associational models* $\mathtt{h_a} : \mathtt{X} \rightarrow \mathtt{Y}$ *that may use all available features, such that* $\mathtt{f} \in \mathcal{H}_\mathtt{C}$ *and* $\mathcal{H}_\mathtt{C} \subseteq \mathcal{H}_\mathtt{A}$*.*

1. *When generation of* $\mathtt{Y}$ *is deterministic,* $\mathtt{y} = \mathtt{f}(\mathtt{X_c})$ *(e.g., when* $\mathtt{Y|X_C}$ *is almost surely constant), the* $\mathtt{ODE}$ *loss for a causal model* $\mathtt{h_c} \in \mathcal{H}_\mathtt{C}$ *is bounded by:*

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c},\mathtt{y}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h_c},\mathtt{y}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h_c},\mathtt{y}) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{C}}(\mathtt{P},\mathtt{P}^*) + \mathtt{IDE}_\mathtt{P}(\mathtt{h_c},\mathtt{y}) \qquad (4)\end{aligned}$$

   Further, for any $\mathtt{P}$ *and* $\mathtt{P}^*$*, the upper bound of* $\mathtt{ODE}$ *from a dataset* $\mathtt{S} \sim \mathtt{P(X,Y)}$ *to* $\mathtt{P}^*$*(called* $\mathtt{ODE}\text{-}\mathtt{Bound}$*) for a causal model* $\mathtt{h_c} \in \mathcal{H}_\mathtt{C}$ *is less than or equal to the upper bound* $\mathtt{ODE}\text{-}\mathtt{Bound}$ *of an associational model* $h_a \in \mathcal{H}_A$*, with probability at least* $(1-\delta)^2$*.*

$$\mathtt{ODE}\text{-}\mathtt{Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c},\mathtt{y};\delta) \leq \mathtt{ODE}\text{-}\mathtt{Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a},\mathtt{y};\delta)$$

2. *When generation of* $\mathtt{Y}$ *is probabilistic, the* $\mathtt{ODE}$ *error for a causal model* $\mathtt{h_c} \in \mathcal{H}_\mathtt{C}$ *includes additional terms for the loss between* $\mathtt{Y}$ *and optimal causal models* $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$ *on* $\mathtt{P}$ *and* $\mathtt{P}^*$ *respectively.*

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c},\mathtt{y}) \leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{C}}(\mathtt{P},\mathtt{P}^*) + \mathtt{IDE}_\mathtt{P}(\mathtt{h_c},\mathtt{y}) +& \\ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*},\mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}},\mathtt{y})& \qquad (5)\end{aligned}$$

   While the loss of an associational model can be lower on $\mathtt{P}$*, there always exists a* $\mathtt{P}^*$ *such that the worst case* $\mathtt{ODE}\text{-}\mathtt{Bound}$ *for an associational model is higher than the same for a causal model.*

$$\max_{\mathtt{P}^*} \mathtt{ODE}\text{-}\mathtt{Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c},\mathtt{y};\delta) \leq \max_{\mathtt{P}^*} \mathtt{ODE}\text{-}\mathtt{Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a},\mathtt{y};\delta)$$

Proof. The proof has three parts: General ODE Bound for a model, equivalence of loss-minimizing causal models on $\mathtt{P}$ and $\mathtt{P}^*$, and finally the two claims from the Theorem.

I. GENERAL ODE BOUND
Consider a model $\mathtt{h} : \mathtt{X} \rightarrow \mathtt{Y}$ belonging to a set of models $\mathcal{H}$, that was trained on $\mathtt{S} \sim \mathtt{P(X,Y)}$. From Def. 2 we write,

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h},\mathtt{y}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h},\mathtt{y}) \\ &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) + \\ &\quad \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h},\mathtt{y}) \\ &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) + \mathtt{IDE}_\mathtt{P}(\mathtt{h},\mathtt{y})\end{aligned} \qquad (8)$$

where the last equation is to due to Def.1 of the in-distribution generalization error.

Let us denote the optimal loss-minimizing hypotheses over $\mathcal{H}$ for $\mathtt{P}$ and $\mathtt{P}^*$ as $\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}$ and $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*}$.

$$\mathtt{h}^{\mathtt{OPT}}_\mathtt{P} = \arg\min_{\mathtt{h}\in\mathcal{H}} \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) \qquad \mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*} = \arg\min_{\mathtt{h}\in\mathcal{H}} \mathcal{L}_{\mathtt{P}*}(\mathtt{h},\mathtt{y}) \qquad (9)$$

Using the triangle inequality of the loss function, we can write:

$$\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y}) \leq \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) \qquad (10)$$

And,

$$\begin{aligned}\mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) &\geq \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) - \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) \\ \Rightarrow -\mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y}) &\leq -\mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y})\end{aligned} \qquad (11)$$

Thus, combining Eqns. 8, 10 and 11, we obtain,

$$\begin{aligned}&\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h},\mathtt{y}) \\ &\quad \leq \mathtt{IDE}_\mathtt{P}(\mathtt{h},\mathtt{y}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) + \\ &\quad \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) \\ &\quad = \mathtt{IDE}_\mathtt{P}(\mathtt{h},\mathtt{y}) + (\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}) - \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_\mathtt{P})) + \\ &\quad \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{f}) \\ &\quad \leq \mathtt{IDE}_\mathtt{P}(\mathtt{h},\mathtt{y}) + \mathtt{disc}_{\mathtt{L},\mathcal{H}}(\mathtt{P},\mathtt{P}^*) + \\ &\quad \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P},\mathtt{y})\end{aligned} \qquad (12)$$

where the last inequality is due to the definition of discrepancy distance (Definition 3).

Below we show that Eqn. 12 divides the out-of-distribution generalization error of a model $\mathtt{h}$ in four parts. As defined in the Theorem statement, $\mathcal{H}_C$ refers to the class of models that uses all causal features ($\mathtt{X_C}$), parents of $\mathtt{Y}$ over the structural causal graph; and $\mathcal{H}_\mathcal{A}$ refers to the class of associational models that may use all or a subset of all available features.

1. $\mathtt{IDE_P(h,y)}$ denotes the in-distribution error of $\mathtt{h}$. This can be bounded by typical generalization bounds, such as the uniform error bound that depends only on the VC dimension and sample size of $\mathtt{S}$ (Shalev-Shwartz & Ben-David, 2014). Using a uniform error bound based on the VC dimension, we obtain, with probability at least $1-\delta$,

$$\begin{aligned}\mathtt{IDE} &\leq \sqrt{8\frac{\mathtt{VCdim}(\mathcal{H})(\ln(2|\mathtt{S}|)+1)+\ln(4/\delta)}{|\mathtt{S}|}}\\ &= \mathtt{IDE\text{-}Bound}(\mathcal{H},\mathtt{S})\end{aligned} \tag{13}$$

   Since $\mathcal{H}_\mathtt{C} \subseteq \mathcal{H}_\mathtt{A}$, VC-dimension of causal models is not greater than that of associational models. Thus,

$$\begin{aligned}\mathtt{VCDim}(\mathcal{H}_\mathtt{C}) \leq \mathtt{VCDim}(\mathcal{H}_\mathtt{A}) \Rightarrow \mathtt{IDE\text{-}Bound}(\mathcal{H}_\mathtt{C},\mathcal{S})\\ \leq \mathtt{IDE\text{-}Bound}(\mathcal{H}_\mathtt{A},\mathcal{S})\end{aligned} \tag{14}$$

2. $\mathtt{disc}_{\mathtt{L},\mathcal{H}}(\mathtt{P},\mathtt{P}^*)$ denotes the distance between the two distributions. Given two distributions, the discrepancy distance does not depend on $\mathtt{h}$, but only on the model class $\mathcal{H}$. From Definition 3, discrepancy distance is the maximum quantity over all pairs of models in a model class. Since $\mathcal{H}_\mathtt{C} \subseteq \mathcal{H}_\mathtt{A}$, we obtain that:

$$\mathtt{disc}_{\mathtt{L},\mathcal{H}_C}(\mathtt{P},\mathtt{P}^*) \leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{A}}(\mathtt{P},\mathtt{P}^*) \tag{15}$$

3. $\mathcal{L}_\mathtt{P}(\mathtt{h}_\mathtt{P}^\mathtt{OPT},\mathtt{y})$ measures the error of the loss-minimizing model on $\mathtt{P}$, when evaluated on $\mathtt{P}$. While $\mathtt{h}_\mathtt{P}^\mathtt{OPT}$ is optimal, there can still be error due to the true labeling function $\mathtt{f}$ being outside the model class $\mathcal{H}$, or irreducible error due to probabilistic generation of Y.
4. $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_\mathtt{P}^\mathtt{OPT},\mathtt{y})$ measures the error of the loss-minimizing model on $\mathtt{P}$, when evaluated on $\mathtt{P}^*$. In addition to the reasons cited above, this error can be due to differences in both $\Pr(\mathtt{X})$ and $\Pr(\mathtt{Y}|\mathtt{X})$ between $\mathtt{P}$ and $\mathtt{P}^*$: change in the marginal distribution of inputs $\mathtt{X}$, and/or change in the conditional distribution of $\mathtt{Y}$ given $\mathtt{X}$.

II. SAME LOSS-MINIMIZING CAUSAL MODEL OVER $\mathtt{P}$ AND $\mathtt{P}^*$

Below we show that for a given distribution $\mathtt{P}$ and another distribution $\mathtt{P}^*$ such that $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$, the loss minimizing model is the same for causal models ($\mathtt{h}_\mathtt{c,P}^\mathtt{OPT} = \mathtt{h}_{\mathtt{c,P}^*}^\mathtt{OPT}$), but not necessarily for associational models.

Causal Model. Given a structural causal network, let us construct a model using all parents of $\mathtt{X_C}$ of $\mathtt{Y}$. By property of the structural causal network, $\mathtt{X_C}$ includes all parents of $\mathtt{Y}$ and therefore there are no backdoor paths. Using Rule 2 of do-calculus from Pearl (2009):

$$\Pr(\mathtt{Y}|\mathtt{do}(\mathtt{X_c}=\mathtt{x_c})) = \mathtt{P}(\mathtt{Y}|\mathtt{X_C}=\mathtt{x_c}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C}=\mathtt{x_c}) \tag{16}$$

where the last equality is assumed since data from $\mathtt{P}^*$ also shares the same causal graph. Defining $\mathtt{h}_\mathtt{c,P}^\mathtt{OPT} = \arg\min_{\mathtt{h_c}\in\mathcal{H}_C} \mathcal{L}_\mathtt{P}(\mathtt{h_c},\mathtt{y})$ and $\mathtt{h}_{\mathtt{c,P}^*}^\mathtt{OPT} = \arg\min_{\mathtt{h_c}\in\mathcal{H}_C} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h_c},\mathtt{y})$, we can write,

$$\begin{aligned}\mathtt{h}_\mathtt{c,P}^\mathtt{OPT} &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{C}} \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y})\\ &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{C}} \mathbb{E}_{\mathtt{P}(\mathtt{x_c},\mathtt{y})}\mathtt{L}(\mathtt{h}(\mathtt{x_c}),\mathtt{y}) = \mathtt{f}_{\mathtt{P}(\mathtt{Y}|\mathtt{X_C})}\end{aligned} \tag{17}$$

since $\mathtt{f} = \arg\min_\mathtt{h} \mathtt{L_x}(\mathtt{h}(\mathtt{x_c}),\mathtt{y})$ for all $\mathtt{x_c}$ and thus does not depend on $\Pr(\mathtt{X_C})$, and $\mathtt{f}\in\mathcal{H}_\mathtt{C}$. Similarly, for $\mathtt{h}_{\mathtt{c,P}^*}^\mathtt{OPT}$, we can write:

$$\begin{aligned}\mathtt{h}_{\mathtt{c,P}^*}^\mathtt{OPT} &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{C}} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y})\\ &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{C}} \mathbb{E}_{\mathtt{P}^*(\mathtt{x_c},\mathtt{y})}\mathtt{L}(\mathtt{h}(\mathtt{x_c}),\mathtt{y}) = \mathtt{f}_{\mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})}\end{aligned} \tag{18}$$

Since $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$, we obtain,

$$\mathtt{f}_{\mathtt{P}(\mathtt{Y}|\mathtt{X_C})} = \mathtt{f}_{\mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})} \Rightarrow \mathtt{h}_\mathtt{c,P}^\mathtt{OPT} = \mathtt{h}_{\mathtt{c,P}^*}^\mathtt{OPT} \tag{19}$$

Associational Model. In contrast, an associational model may use a subset $\mathtt{X_A}\subseteq\mathtt{X}$ that may not include all parents of $\mathtt{Y}$, or may include parents but also include other extraneous variables. Following the derivation for causal models, let us define $\mathtt{h}_\mathtt{a,P}^\mathtt{OPT} = \arg\min_{\mathtt{h_a}\in\mathcal{H}_\mathtt{A}} \mathcal{L}_\mathtt{P}(\mathtt{h_a},\mathtt{y})$ and $\mathtt{h}_{\mathtt{a,P}^*}^\mathtt{OPT} = \arg\min_{\mathtt{h_a}\in\mathcal{H}_\mathcal{A}} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h_a},\mathtt{y})$, we can write,

$$\begin{aligned}\mathtt{h}_\mathtt{a,P}^\mathtt{OPT} &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{A}} \mathcal{L}_\mathtt{P}(\mathtt{h},\mathtt{y})\\ &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{A}} \mathbb{E}_{\mathtt{P}(\mathtt{x_a},\mathtt{y})}\mathtt{L}(\mathtt{h}(\mathtt{x_a}),\mathtt{y}) = \mathtt{f}_{\mathtt{P}(\mathtt{X_A},\mathtt{Y})}\end{aligned} \tag{20}$$

where we define $\mathtt{f_A}$ as, $\mathtt{f_A} = \arg\min_\mathtt{h} \mathtt{L_x}(\mathtt{h}(\mathtt{x_a}),\mathtt{y})$ for any $\mathtt{x_a}$. Similarly, for $\mathtt{h}_{\mathtt{a,P}^*}^\mathtt{OPT}$, we can write:

$$\begin{aligned}\mathtt{h}_{\mathtt{a,P}^*}^\mathtt{OPT} &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{A}} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y})\\ &= \arg\min_{\mathtt{h}\in\mathcal{H}_\mathtt{A}} \mathbb{E}_{\mathtt{P}^*(\mathtt{x_a},\mathtt{y})}\mathtt{L}(\mathtt{h}(\mathtt{x_a}),\mathtt{y}) = \mathtt{f}_{\mathtt{P}^*(\mathtt{X_A},\mathtt{Y})}\end{aligned} \tag{21}$$

Now, in general,

$$\mathtt{P}(\mathtt{X_A},\mathtt{Y}) \neq \mathtt{P}^*(\mathtt{X_A},\mathtt{Y}) \Rightarrow \mathtt{f}_{\mathtt{P}(\mathtt{X_A},\mathtt{Y})} \neq \mathtt{f}_{\mathtt{P}^*(\mathtt{X_A},\mathtt{Y})}$$

Even if the optimal associational model $\mathtt{f_A}\in\mathcal{H}_\mathtt{A}$ (as we assumed for causal models), and thus $\mathtt{f}_{\mathtt{P}(\mathtt{X_A},\mathtt{Y})} = \mathtt{f}_{\mathtt{P}(\mathtt{Y}|\mathtt{X_A})}$ and

$\mathtt{f}_{\mathtt{P}^*(\mathtt{X}_\mathtt{A},\mathtt{Y})} = \mathtt{f}_{\mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{A})}$, they are not the same since $\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{A}) \neq \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{A})$. Therefore we obtain,

$$\mathtt{f}_{\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{A})} \neq \mathtt{f}_{\mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{A})} \Rightarrow \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*} \tag{22}$$

That said, since $\mathtt{X}_\mathtt{C} \subset \mathtt{X}$, it is possible that $\mathtt{X}_\mathtt{A} = \mathtt{X}_\mathtt{C}$ for some $\mathtt{X}$ and $\mathcal{H}$, and thus the loss-minimizing associational model includes only the causal features of $\mathtt{Y}$. Then $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$. In general, though, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$.

IIIa. CLAIM 1

As a warmup, consider the case when $\mathtt{Y}$ is generated deterministically. That is, the optimal model $\mathtt{f}$ has zero error. Then, both the loss-minimizing causal model and loss-minimizing associational model have zero error when evaluated on the same distribution that they were trained on. Thus, $\mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) = \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) = 0$. Similarly, $\mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) = 0$. (Note that here we consider only those cases where $\mathtt{f}_{\mathtt{P}(\mathtt{Y}|\mathtt{X})} \in \mathcal{H}_\mathtt{A}$ and $\mathtt{f}_{\mathtt{P}^*(\mathtt{Y}|\mathtt{X})} \in \mathcal{H}_\mathtt{A}$ for a fair comparison; otherwise, the error bound for $\mathtt{h}_\mathtt{a} \in \mathcal{H}_\mathtt{A}$ is trivially larger than that for $\mathtt{h}_\mathtt{c} \in \mathcal{H}_\mathtt{C}$).

Further, for a causal model, using Equation 19, we obtain:

$$\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) = \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) = 0 \tag{23}$$

However, the same does not hold for associational models: $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y})$ need not be zero.

We now present the loss bounds. Using Equations 19 and 23, we write Equation 12 for a causal model as:

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h}_\mathtt{c}, \mathtt{y}) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE}_\mathtt{P}(\mathtt{h}_\mathtt{c}, \mathtt{y})\end{aligned} \tag{24}$$

For an associational model, we obtain,

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h}_\mathtt{a}, \mathtt{y}) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{A}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE}_\mathtt{P}(\mathtt{h}_\mathtt{a}, \mathtt{y}) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y})\end{aligned} \tag{25}$$

Using Eqn. 13 that bounds IDE with probability $1 - \delta$, and Eqns. 14 and 15 that compare IDE-Bound and discrepancy distance between causal and associational model classes, we can rewrite Eqn. 24. With probability at least $1 - \delta$:

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}) &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{C}, \mathtt{S}; \delta) \\ &= \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{A}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{A}, \mathtt{S}; \delta)\end{aligned} \tag{26}$$

Similarly, for the associational model,

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}) &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{A}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{A}, \mathtt{S}; \delta) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) \\ &= \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}; \delta)\end{aligned} \tag{27}$$

Therefore, comparing Eqn. 26 and 27, we claim for any $\mathtt{P}$ and $\mathtt{P}^*$, with probability $(1 - \delta)^2$,

$$\mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \leq \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}; \delta) \tag{28}$$

IIIb. CLAIM 2

We now consider the general case when $\mathtt{Y}$ is generated probabilistically. Thus, even though $\mathtt{f} \in \mathcal{H}_\mathtt{C}$ and $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*} = \mathtt{f}$, $\mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) \neq 0$ and $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) \neq 0$.

Using the IDE bound from Eqn. 13, we write Eqn. 12 as,

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}) &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE}_\mathtt{P}(\mathtt{h}_\mathtt{c}, \mathtt{y}) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{C}, \mathtt{S}; \delta) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) \qquad (29) \\ &= \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \\ &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{A}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{A}, \mathtt{S}; \delta) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) \qquad (30)\end{aligned}$$

where Eqn. 29 uses $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$ and Eqn. 30 uses inequalities comparing IDE and discrepancy distance from Eqns. 14 and 15.

Similarly, for associational model,

$$\begin{aligned}\mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}) &= \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{A}}(\mathtt{P}, \mathtt{P}^*) \\ &+ \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{A}, \mathtt{S}; \delta) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y})\end{aligned} \tag{31}$$

Now, we compare the last two terms of Equations 30 and 31. Since $\mathcal{H}_\mathcal{C} \subseteq \mathcal{H}_\mathcal{A}$, loss of the loss-minimizing associational model can be lower than the loss of the causal model trained on the same distribution. Thus, $\mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) \leq \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y})$.

However, since $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$, loss of the loss-minimizing associational model trained on $\mathtt{P}$ can be higher on $\mathtt{P}^*$ than the loss of optimal causal model trained on $\mathtt{P}^*$ and evaluated on $\mathtt{P}^*$. Formally, let $\gamma_1 \geq 0$ be the loss reduction over $\mathtt{P}$ due to use of associational model optimized on $\mathtt{P}$, compared to the loss-minimizing causal model. Similarly, let γ_2 be the increase in loss over $\mathtt{P}^*$ due to using the associational model optimized over $\mathtt{P}$, compared to the loss-minimizing causal model.

$$\gamma_1 = \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) \tag{32}$$

$$\gamma_2 = \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) - \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) \tag{33}$$

Then, Eqn. 31 transforms to,

$$\begin{aligned}\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}) &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathcal{A}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_\mathtt{P}(\mathcal{H}_\mathtt{A}, \mathcal{S}; \delta) \\ &+ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) + \gamma_2 - \gamma_1\end{aligned} \tag{34}$$

Hence, as long as $\gamma_2 \geq \gamma_1$, we obtain,

$$\mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \leq \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}; \delta) \tag{35}$$

Below we show that such a $\mathtt{P}^*$ always exists, and further, the worst-case $\max_{\mathtt{P}^*} \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}, \mathtt{y}; \delta)$ is always lower for a causal model than an associational model.

There exists $\mathtt{P}^*$ such that $\gamma_2 \geq \gamma_1$. The proof is by construction. As an example, consider L1 loss and a distribution $\mathtt{P}$ such that the optimal causal model f for an input data point $\mathtt{x}^{(\mathtt{i})}$ can be written as,

$$\mathtt{y}^{(\mathtt{i})} = \mathtt{f}_\mathtt{P}(\mathtt{x}_\mathtt{C}^{(\mathtt{i})}) + \xi_\mathtt{i} = \mathtt{f}_{\mathtt{P}^*}(\mathtt{x}_\mathtt{C}^{(\mathtt{i})}) + \xi_\mathtt{i} \tag{36}$$

where $\mathtt{f}(\mathtt{x}_\mathtt{C}) = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$ refers to the optimal causal model and is the same for $\mathtt{P}$ and $\mathtt{P}^*$ (using Eqn. 19). Let $\mathtt{f}_\mathtt{P}(\mathtt{x}_\mathtt{A}) = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}$ be the optimal associational model over $\mathtt{P}$. We can rewrite $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}$ as an arbitrary change from $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}$, using $\lambda^{(\mathtt{i})}_{\mathtt{x}_\mathtt{A}}$ as a parameter that can be different for each data point $\mathtt{x}^{(\mathtt{i})}$. That is,

$$\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}(\mathtt{x}^{(\mathtt{i})}) = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}(\mathtt{x}_\mathtt{C}^{(\mathtt{i})}) + \lambda^{(\mathtt{i})}_{\mathtt{x}_\mathtt{A}} \tag{37}$$

Based on Eqns. 36 and 37, γ_1 can be written as,

$$\begin{aligned} \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) &= \mathbb{E}_\mathtt{P}[|\xi|] \\ \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) &= \mathbb{E}_\mathtt{P}[|\xi - \lambda_{\mathtt{x}_\mathtt{A}}|] \\ \Rightarrow \gamma_1 &= \mathbb{E}_\mathtt{P}[|\lambda_{\mathtt{x}_\mathtt{A}}|] \end{aligned} \tag{38}$$

Then, we can construct a $\mathtt{P}^*(\mathtt{X}, \mathtt{Y})$ such that (i) the relationship ($\Pr(\mathtt{Y}|\mathtt{X}_\mathtt{A})$) between $\mathtt{x}_\mathtt{A}$ and $\mathtt{y}$ is reversed, and (ii) $\Pr(\mathtt{X})$ is chosen such that $\mathbb{E}_{\mathtt{P}^*}[\lambda_{\mathtt{x}_\mathtt{A}}] \geq \mathbb{E}_\mathtt{P}[\lambda_{\mathtt{x}_\mathtt{A}}]$ (e.g., by assigning higher probability weights to data points $\mathtt{i}$ where $|\lambda^{(\mathtt{i})}_{\mathtt{x}_\mathtt{A}}|$ is high). That is, consider a $\mathtt{P}^*$ such that we can write $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$ as,

$$\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}(\mathtt{x}^{(\mathtt{i})}) = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}(\mathtt{x}_\mathtt{C}^{(\mathtt{i})}) - \lambda^{(\mathtt{i})}_{\mathtt{x}_\mathtt{A}} \tag{39}$$

On such $\mathtt{P}^*$, the loss-minimizing causal model remains the same. However, the loss of the associational model $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}$ on such $\mathtt{P}^*$ increases and can be written as:

$$\begin{aligned} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) &= \mathbb{E}_{\mathtt{P}^*}[|\xi|] \\ \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}) &= \mathbb{E}_{\mathtt{P}^*}[|\xi + \lambda_{\mathtt{x}_\mathtt{A}}|] \\ \Rightarrow \gamma_2 &= \mathbb{E}_{\mathtt{P}^*}[|\lambda_{\mathtt{x}_\mathtt{A}}|] \end{aligned} \tag{40}$$

From condition (ii) above, $\mathbb{E}_{\mathtt{P}^*}[\lambda_{\mathtt{x}_\mathtt{A}}] \geq \mathbb{E}_\mathtt{P}[\lambda_{\mathtt{x}_\mathtt{A}}]$, thus $\gamma_2 \geq \gamma_1$.

Note that we did not use any special property of the L1 Loss above. In general, we can write the loss-minimizing function $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}$ as adding some arbitrary value $\lambda^{(\mathtt{i})}_{\mathtt{x}_\mathtt{A}}$ to $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}(\mathtt{x}_\mathtt{C}^{(\mathtt{i})})$; and then construct a $\mathtt{P}^*$ such that the relationship $\Pr(\mathtt{Y}|\mathtt{X}_\mathtt{A})$ is reversed on $\mathtt{P}^*$, and thus $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$ subtracts the same value. Further, the input data distribution $\mathtt{P}^*(\mathtt{X})$ can be chosen such that $\gamma_2 \geq \gamma_1$. That is, for a loss $\mathtt{L}$, we can choose λ such that $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}; \lambda) - \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{y}) \geq \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y}; \lambda)$.

Hence, there exists a $\mathtt{P}^*$ such that $\gamma_2 \geq \gamma_1$, and thus,

$$\mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \leq \mathtt{ODE{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}; \delta) \tag{41}$$

Worst case ODE-bound for causal model is lower. Finally, we show that the for a fixed $\mathtt{P}$, the worst case $\mathtt{ODE\text{-}Bound}$ also follows Eqn. 41. Looking at Eqns. 30 and 31, $\mathtt{ODE\text{-}Bound}$ will be highest for a $\mathtt{P}^*$ such that discrepancy between $\mathtt{P}$ and $\mathtt{P}^*$ is highest and $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y})$ is highest. Below we show that discrepancy $\mathtt{disc}_\mathtt{L}(\mathtt{P}, \mathtt{P}^*)$ increases as $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y})$ increases.

$$\begin{aligned} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) \\ &\leq \mathtt{disc}_\mathtt{L}(\mathtt{P}, \mathtt{P}^*) + \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) \\ \Rightarrow \mathtt{disc}_\mathtt{L}(\mathtt{P}, \mathtt{P}^*) &\geq \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) - \mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y}) \end{aligned} \tag{42}$$

where $\mathcal{L}_\mathtt{P}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y})$ is fixed since $\mathtt{P}$ is fixed. Thus, the above equation shows that whenever $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y})$ is high, discrepancy is also high. Hence, for any $\mathtt{P}^*_{\mathtt{max}}$ that maximizes $\mathtt{ODE\text{-}Bound}$, $\mathtt{P}^*_{\mathtt{max}} = \arg\max_{\mathtt{P}^*} \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}, \mathtt{y}; \delta)$, $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_\mathtt{P}, \mathtt{y})$ is also maximized.

Now, let us consider causal and associational models, and their respective worst case $\mathtt{P}^*_{\mathtt{max}}$. To complete the proof, we need to check whether $\gamma_2 \geq \gamma_1$ for such maximal $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y})$ and $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y})$. Since γ_2 increases monotonically with $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{y})$ ($\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y})$ is bounded by $\max_\mathtt{x} \mathtt{L}_\mathtt{x}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{y})$), and there exists at least one $\mathtt{P}^*$ such that $\gamma_2 \geq \gamma_1$, this implies that $\gamma_2 \geq \gamma_1$ for $\mathtt{P}^*_{\mathtt{max}}$ too. Therefore, using Equation 41,

$$\max_{\mathtt{P}^*} \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{y}; \delta) \leq \max_{\mathtt{P}^*} \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{y}; \delta) \tag{43}$$

□

A.2 Generalization over a Single Datapoint

Theorem 2. *Consider a causal model* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} : \mathtt{X}_\mathtt{C} \rightarrow \mathtt{Y}$ *and an associational model* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}} : \mathtt{X} \rightarrow \mathtt{Y}$ *trained on a dataset* $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$ *with loss* L. *Let* $(\mathtt{x}, \mathtt{y}) \in \mathtt{S}$ *and* $(\mathtt{x}', \mathtt{y}') \notin \mathtt{S}$ *be two input instances such that they share the same true labelling function on the causal features,* $\mathtt{y} \sim \mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C} = \mathtt{x})$ *and* $\mathtt{y}' \sim \mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C} = \mathtt{x}')$. *Then, the worst-case generalization error for the causal model on such* $\mathtt{x}'$ *is less than or equal to that for the associational model.*

$$\max_{\mathtt{x} \in \mathtt{S}, \mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}, \mathtt{y}) - \mathtt{L}_\mathtt{x}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}, \mathtt{y}) \leq \max_{\mathtt{x} \in \mathtt{S}, \mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}, \mathtt{y}) - \mathtt{L}_\mathtt{x}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}, \mathtt{y})$$

Proof. For any model $\mathtt{h}$, we can write,

$$\max_{\mathtt{x} \in \mathtt{S}, \mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}, \mathtt{y}) - \mathtt{L}_\mathtt{x}(\mathtt{h}, \mathtt{y}) = \max_{\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}, \mathtt{y}) - \min_{\mathtt{x} \in \mathtt{S}} \mathtt{L}_\mathtt{x}(\mathtt{h}, \mathtt{y}) \tag{44}$$

since $\mathtt{x}'$ and $\mathtt{x}$ are independently selected. To prove the main result, we will show that the maximum loss on an unseen x', $\max_{x'} L_{x'}(h, y)$ is higher for a loss-minimizing associational model than a causal model, and that minimum loss on a training point $x \in S$, $\min_{x \in S} L_x(h, y)$ is lower for the associational model than a causal model.

Loss on a training data point. First, consider loss on $x \in S$, $L_x(h, y)$.

$$h^{min}_{c,S} = \arg\min_{h \in \mathcal{H}_C} \mathcal{L}_S(h_c, y) = \arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} L_{x_i}(h, y)$$

$$h^{min}_{a,S} = \arg\min_{h \in \mathcal{H}_\mathcal{A}} \mathcal{L}_S(h_a, y) = \arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} L_{x_i}(h, y)$$

Since $\mathcal{H}_\mathcal{C} \subseteq \mathcal{H}_\mathcal{A}$, the average training loss will be lower for the associational model.

$$\mathcal{L}_S(h^{min}_{c,S}, y) \geq \mathcal{L}_S(h^{min}_{a,S}, y) \tag{45}$$

Further, under a suitably complex $\mathcal{H}_A$ there exists a $h^{min}_{a,S}$ such that the loss L is lower for any $x \in S$. Therefore,

$$\min_{x \in S} L_x(h^{min}_{c,S}, y) \geq \min_{x \in S} L_x(h^{min}_{a,S}, y) \tag{46}$$

Loss on an unseen data point. Second, consider $L_{x'}(h, y)$. Without loss of generality, let us write the true function for some $(x', y') \sim P^*$ as,

$$y' = h^{OPT}_{c,P^*}(x'_c) + \epsilon = h^{OPT}_{c,P}(x'_c) + +\epsilon \tag{47}$$

where we use that $h^{OPT}_{c,P} = h^{OPT}_{c,P^*}$. Suppose there is a data point (x'_1, y'_1) such that the loss L is maximum for $h^{min}_{c,S}$.

$$\begin{aligned} \max_{x' \notin S} L_{x'}(h^{min}_{c,S}, y) &= L_{x'_1}(h^{min}_{c,S}(x'_1), y'_1) \\ &= L_{x'_1}(h^{min}_{c,S}(x'_{c,1}), h^{OPT}_{c,P}(x'_{c,1}) + \epsilon_1) \end{aligned} \tag{48}$$

Now for the associational model $h^{min}_{a,S}$, the corresponding loss on x'_1 is,

$$L_{x'_1}(h^{min}_{a,S}, y) = L_{x'_1}(h^{min}_{a,S}, h^{OPT}_{c,P} + \epsilon_1) \tag{49}$$

Without loss of generality, we can write the output of the associational model $h^{min}_{a,S}$ on a particular input x' as,

$$h^{min}_{a,S}(x') = h^{min}_{c,S}(x'_c) + h_a(x') \tag{50}$$

where h_a is some associational function of x. Therefore the loss on x'_1 becomes,

$$L_{x'_1}(h^{min}_{a,S}, y) = L_{x'_1}(h^{min}_{c,S} + h_a, h^{OPT}_{c,P} + \epsilon_1) \tag{51}$$

Since $\Pr(Y|X_A)$ can change for different $x' \sim P^*$ (where $X_A = X \setminus X_C$ refers to the associational features), we will show that RHS of Eqn. 49 can always be greater than or equal to the RHS of Eqn. 48. For ease of exposition, we consider L1 loss below. For a causal model, the loss can be written as,

$$\begin{aligned} &L_{x'_1}(h^{min}_{c,S}, h^{OPT}_{c,P} + +\epsilon) \\ &\qquad = |h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1}) - \epsilon_1| \\ &\qquad = |h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1})| + |\epsilon_1| \end{aligned} \tag{52}$$

where x'_1 (and thus ϵ_1) is chosen such that $\epsilon_1(h^{OPT}_{c,P}(x'_{c,1}) - h^{min}_{c,S}(x'_{c,1})) \geq 0$ which leads to maximum loss. And for the associational model, the loss on the same (x'_1, y'_1) can be written as,

$$\begin{aligned} &L_{x'_1}(h^{min}_{a,S}, h^{OPT}_{c,P} + \epsilon_1) \\ &\quad = |h^{min}_{c,S}(x'_{c,1}) + h_a(x'_1) - h^{OPT}_{c,P}(x'_{c,1}) - \epsilon_1| \\ &\quad = |(h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1})) + (h_a(x'_1) - \epsilon_1)| \end{aligned} \tag{53}$$

Comparing Eqns. 52 and 53, two cases arise. If $h_a(x'_1)\epsilon_1 \leq 0$, then we obtain,

$$\begin{aligned} &L_{x'_1}(h^{min}_{a,S}, h^{OPT}_{c,P} + \epsilon_1) \\ &\quad = |h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1})| + |(h_a(x'_1) - \epsilon_1)| \\ &\quad = |h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1})| + |h_a(x'_1)| + |\epsilon_1| \end{aligned} \tag{54}$$

which is greater than maximum loss on x'_1 using a causal model (Eqn. 52). Otherwise, if $h_a(x'_1)\epsilon_1 > 0$, we can sample a new data point (x'_2, y'_2) from some other P^* such that its causal features are the same $(x'_{c,1} = x'_{c,2})$ and thus y is the same $(y'_2 = y'_1 = h^{OPT}_{c,P}(x'_{c,1}) + \epsilon_1)$, but its associational features are different $(x'_{a,1} \neq x'_{a,2})$. Specifically, $x'_{a,2}$ is chosen such that $h_a(x'_2)\epsilon_1 \leq 0$. Thus we again obtain,

$$\begin{aligned} &L_{x'_2}(h^{min}_{a,S}, h^{OPT}_{c,P} + \epsilon_1) \\ &\quad = |(h^{min}_{c,S}(x'_{c,2}) - h^{OPT}_{c,P}(x'_{c,2})) + (h_a(x'_2) - \epsilon_1)| \\ &\quad = |(h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1})) + (h_a(x'_2) - \epsilon_1)| \\ &\quad = |(h^{min}_{c,S}(x'_{c,1}) - h^{OPT}_{c,P}(x'_{c,1}))| + |(h_a(x'_2)| + |\epsilon_1| \end{aligned} \tag{55}$$

where the second equality uses $x'_{c,2} = x'_{c,1}$. Combining Eqns. 54 and 55 and comparing to Eqn. 52, we obtain,

$$\max_{x'} L_{x'}(h^{min}_{c,S}, y) \leq \max_{x'} L_{x'}(h^{min}_{a,S}, y) \tag{56}$$

Finally, using Eqns. 46 and 56 leads to the main result.

$$\begin{aligned} &\max_{x'} L_{x'}(h^{min}_{c,S}, y) - \min_{x \in S} L_x(h^{min}_{c,S}, y) \\ &\leq \max_{x'} L_{x'}(h^{min}_{a,S}, y) - \min_{x \in S} L_x(h^{min}_{a,S}, y) \\ &\max_{x', x \in S} L_{x'}(h^{min}_{c,S}, y) - L_x(h^{min}_{c,S}, y) \leq \max_{x', x \in S} L_{x'}(h^{min}_{a,S}, y) - L_x(h^{min}_{a,S}, y) \end{aligned} \tag{57}$$

Using Eqns. 45 and 56 we also obtain an auxiliary result.

$$\max_{x'} L_{x'}(h^{min}_{c,S}, y) - \mathcal{L}_S(h^{min}_{c,S}, y) \leq \max_{x'} L_{x'}(h^{min}_{a,S}, y) - \mathcal{L}_S(h^{min}_{a,S}, y) \tag{58}$$

□

B Sensitivity of Causal and Associational Models

Before we prove Lemma 1, we prove Corollary 1 and restate a Lemma from Wu et al. (2015) for completeness.

Corollary 1. *Let* $\mathtt{S}$ *be a dataset of* $\mathtt{n}$ $(\mathtt{x},\mathtt{y})$ *values, such that* $\mathtt{y}^{(\mathtt{i})} \sim \mathtt{P}(\mathtt{Y}|\mathtt{X_C} = \mathtt{x}^{(\mathtt{i})}) \forall (\mathtt{x}^{(\mathtt{i})}, \mathtt{y}^{(\mathtt{i})}) \in \mathtt{S}$, *where* $\mathtt{P}(\mathtt{Y}|\mathtt{X_C})$ *is the invariant conditional distribution on the causal features* $\mathtt{X_C}$. *Consider a neighboring dataset* $\mathtt{S}'$ *such that* $\mathtt{S}' = \mathtt{S}\backslash(\mathtt{x},\mathtt{y}) + (\mathtt{x}',\mathtt{y}')$ *where* $(\mathtt{x},\mathtt{y}) \in \mathtt{S}$, $(\mathtt{x}',\mathtt{y}') \notin \mathtt{S}$, *and* $(\mathtt{x}',\mathtt{y}')$ *shares the same conditional distribution* $\mathtt{y}' \sim \mathtt{P}(\mathtt{Y}|\mathtt{X_C} = \mathtt{x}'_c)$. *Then the maximum generalization error from* $\mathtt{S}$ *to* $\mathtt{S}'$ *for a causal model trained on* $\mathtt{S}$ *is lower than or equal to that of an associational model.*

$$\max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) \le \max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})$$

Proof. Let $\mathtt{S_{n-1}} = \mathtt{S} \setminus (\mathtt{x},\mathtt{y}))$ and similarly $\mathtt{S}'_{\mathtt{n-1}} = \mathtt{S}' \setminus (\mathtt{x}',\mathtt{y}')$. Since $\mathtt{S}$ and $\mathtt{S}'$ differ in only one data point, $S_{n-1} = S'_{n-1}$. We will add and subtract sum of losses on data points in $\mathtt{S_{n-1}}$, $(\mathtt{n}-1)\mathtt{L}_{\mathtt{S_{n-1}}}$ to Theorem 2 statement.

Considering the LHS of Theorem 2,

$$\begin{aligned}
&\max_{\mathtt{x}\in\mathtt{S},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&= \max_{\mathtt{x}\in\mathtt{S},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) + (\mathtt{n}-1)\mathcal{L}_{\mathtt{S_{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&- \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - (\mathtt{n}-1)\mathcal{L}_{\mathtt{S_{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&= \max_{\mathtt{x}\in\mathtt{S},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) + (\mathtt{n}-1)\mathcal{L}_{\mathtt{S}'_{\mathtt{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&- \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - (\mathtt{n}-1)\mathcal{L}_{\mathtt{S_{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&= \max_{\mathtt{S}'} \mathtt{n}\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathtt{n}\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})
\end{aligned} \tag{59}$$

Similarly, the RHS of Theorem 2 can be written as,

$$\begin{aligned}
&\max_{\mathtt{x}\in\mathtt{S},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
&= \max_{\mathtt{x}\in\mathtt{S},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) + (\mathtt{n}-1)\mathcal{L}_{\mathtt{S_{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
&- \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - (\mathtt{n}-1)\mathcal{L}_{\mathtt{S_{n-1}}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
&= \max_{\mathtt{S}'} \mathtt{n}\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathtt{n}\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})
\end{aligned} \tag{60}$$

Using Theorem 2 and dividing Eqns. 59 and 60 by $\mathtt{n}$, we obtain,

$$\begin{aligned}
&\max_{\mathtt{S}'} \mathtt{n}\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathtt{n}\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&\le \max_{\mathtt{S}'} \mathtt{n}\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathtt{n}\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
\Rightarrow &\max_{\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&\le \max_{\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})
\end{aligned} \tag{61}$$

Finally, since the above holds for any $\mathtt{S} \sim \mathtt{P}$, it will also hold for the worst-case $\mathtt{S}$. The result follows.

$$\begin{aligned}
&\max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&\le \max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})
\end{aligned} \tag{62}$$

□

Lemma 3. *[From Wu et al. (2015)] Let* $\mathtt{S}$ *and* $\mathtt{S}'$ *be two neighboring datasets as defined in Corollary 1 where* $\mathtt{S}' = \mathtt{S} \setminus (\mathtt{x},\mathtt{y}) + (\mathtt{x}',\mathtt{y}')$. *Given a model class* $\mathcal{H}$, *Let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}}$ *be the loss-minimizing model on* $\mathtt{S}$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'}$ *be the loss-minimizing model on* $\mathtt{S}'$. *Then the difference in losses between the two models on the same dataset is bounded by,*

$$\begin{aligned}
&\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})\\
&\le \frac{\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y})}{\mathtt{n}} + \frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})}{\mathtt{n}}
\end{aligned} \tag{63}$$

Proof. The proof follows from expanding loss over a dataset into individual terms for each data point and then using the fact that $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'}$ has the minimum loss on $\mathtt{S}'$.

Using the definition of $\mathcal{L}_{\mathtt{S}} = \frac{1}{\mathtt{n}}\sum_{\mathtt{i}=1}^{\mathtt{n}} \mathtt{L}_{\mathtt{x_i}}(\mathtt{h},\mathtt{y})$, we can write the following for any two neighboring datasets $\mathtt{S}$ and $\mathtt{S}'$.

$$\begin{aligned}
&\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})\\
&\quad= \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) + \frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y})}{\mathtt{n}}\\
&\quad- (\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) + \frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})}{\mathtt{n}})\\
&\quad= (\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})) + \frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y})}{\mathtt{n}}\\
&\quad+ \frac{\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})}{\mathtt{n}})\\
&\quad\le \frac{\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) - \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y})}{\mathtt{n}} + \frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})}{\mathtt{n}}
\end{aligned} \tag{64}$$

where the last inequality is since $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'}$ is the minimizer of $\mathtt{L}_{\mathtt{S}'}(\mathtt{h},\mathtt{y})$ and thus $\mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'},\mathtt{y}) - \mathcal{L}_{\mathtt{S}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y}) \le 0$.

□

Lemma 1. *Let* $\mathtt{S}$ *and* $\mathtt{S}'$ *be two datasets defined as in Corollary 1. Let a model* $\mathtt{h}$ *be specified by a set of parameters* $\theta \in \Omega \subseteq \mathbb{R}^d$. *Let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}}(\mathtt{x};\theta_{\mathtt{S}})$ *be a model learnt using* $\mathtt{S}$ *as training data and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{S}'}(\mathtt{x};\theta_{\mathtt{S}'})$ *be the model learnt using* $\mathtt{S}'$ *as training data, using a loss function* $\mathtt{L}$ *that is* λ*-strongly convex over* Ω, ρ*-Lipschitz, symmetric and obeys the triangle inequality. Then, under the conditions of Theorem 1 (optimal predictor* $f \in \mathcal{H}_C$*) and for a sufficiently large* $\mathtt{n}$, *the sensitivity of a causal learning function* $\mathcal{F}_c$ *that outputs learnt empirical model* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} \leftarrow \mathcal{F}_{\mathtt{c}}(\mathtt{S})$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}'} \leftarrow \mathcal{F}_{\mathtt{c}}(\mathtt{S}')$ *is lower than or equal to the sensitivity of an associational learning function* $\mathcal{F}_a$ *that outputs* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}} \leftarrow \mathcal{F}_{\mathtt{a}}(\mathtt{S})$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'} \leftarrow \mathcal{F}_{\mathtt{a}}(\mathtt{S}')$,

$$\Delta\mathcal{F}_{\mathtt{c}} = \max_{\mathtt{S},\mathtt{S}'} ||\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}'}||_1 \le \max_{\mathtt{S},\mathtt{S}'} ||\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'}||_1 = \Delta\mathcal{F}_{\mathtt{a}}$$

where the maximum is over all such datasets $\mathtt{S}$ *and* $\mathtt{S}'$.

Proof. Since $\mathtt{L}$ is a strongly convex function over Ω, we can write for the two models $h^{min}_{c,S}$ and $h^{min}_{c,S'}$ trained on S and S' respectively (Wu et al., 2015),

$$\begin{aligned}\mathcal{L}_S(h^{min}_{c,S}, y) &\leq \mathcal{L}_S(\alpha h^{min}_{c,S} + (1-\alpha)h^{min}_{c,S'}, y)\\ &\leq \alpha\mathcal{L}_S(h^{min}_{c,S}, y) + (1-\alpha)\mathcal{L}_S(h^{min}_{c,S'}, y)\\ &\quad - \frac{\lambda}{2}\alpha(1-\alpha)\|h^{min}_{c,S'} - h^{min}_{c,S}\|^2\end{aligned} \tag{65}$$

where $\alpha \in (0, 1)$ and the first inequality is since $h^{min}_{c,S}$ is the loss-minimizing model over S. Rearranging the terms and tending α to 1 leads to,

$$\begin{aligned}&(1-\alpha)(\mathcal{L}_S(h^{min}_{c,S}, y) - \mathcal{L}_S(h^{min}_{c,S'}, y))\\ &\quad\leq -\frac{\lambda}{2}(1-\alpha)\left\|h^{min}_{c,S'} - h^{min}_{c,S}\right\|^2\\ \Rightarrow &\frac{\lambda}{2}\left\|h^{min}_{c,S'} - h^{min}_{c,S}\right\|^2 \leq \mathcal{L}_S(h^{min}_{c,S'}, y) - \mathcal{L}_S(h^{min}_{c,S}, y)\end{aligned} \tag{66}$$

Now consider $\max_{S,S'} \left\|h^{min}_{c,S} - h^{min}_{c,S'}\right\|_1$. Without loss of generality, we can order the pair of datasets S, S' such that $\mathcal{L}_S(h^{min}_{c,S'}, y) \leq \mathcal{L}_{S'}(h^{min}_{c,S}, y)$. Then using Eqn. 66 and taking the maximum, we obtain,

$$\begin{aligned}\frac{\lambda}{2}\max_{S,S'}\left\|h^{min}_{c,S} - h^{min}_{c,S'}\right\|_1^2 &\leq \max_{S,S'}\mathcal{L}_S(h^{min}_{c,S'}, y) - \mathcal{L}_S(h^{min}_{c,S}, y)\\ &\leq \max_{S,S'}\mathcal{L}_{S'}(h^{min}_{c,S}, y) - \mathcal{L}_S(h^{min}_{c,S}, y)\\ &\leq \max_{S,S'}\mathcal{L}_{S'}(h^{min}_{a,S}, y) - \mathcal{L}_S(h^{min}_{a,S}, y)\end{aligned} \tag{67}$$

where the last inequality is due to Theorem 2. Let S_1 and S_1' denote the datasets that lead to the maximum in the RHS above. We know that $\mathcal{L}_{S_1}(h^{min}_{a,S_1'}) \geq \mathcal{L}_{S_1'}(h^{min}_{a,S_1'})$ since $h^{min}_{a,S_1'}$ is the loss-minimizing model over S_1'. Therefore, we can rewrite,

$$\begin{aligned}&\max_{S,S'}\mathcal{L}_{S'}(h^{min}_{a,S}, y) - \mathcal{L}_S(h^{min}_{a,S}, y)\\ &= \mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y)\\ &\leq \mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y) + (\mathcal{L}_{S_1}(h^{min}_{a,S_1'}, y) - \mathcal{L}_{S_1'}(h^{min}_{a,S_1'}, y))\\ &= (\mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1'}(h^{min}_{a,S_1'}, y)) + (\mathcal{L}_{S_1}(h^{min}_{a,S_1'}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y))\end{aligned} \tag{68}$$

Now using Lemma 3, we obtain the following two bounds.

$$\begin{aligned}&\mathcal{L}_{S_1}(h^{min}_{a,S_1'}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y)\\ &\leq \frac{L_{x'}(h^{min}_{a,S_1}, y) - L_{x'}(h^{min}_{a,S_1'}, y)}{n} + \frac{L_x(h^{min}_{a,S_1'}, y) - L_x(h^{min}_{a,S_1}, y)}{n}\\ &\mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1'}(h^{min}_{a,S_1'}, y)\\ &\leq \frac{L_{x'}(h^{min}_{a,S_1}, y) - L_{x'}(h^{min}_{a,S_1'}, y)}{n} + \frac{L_x(h^{min}_{a,S_1'}, y) - L_x(h^{min}_{a,S_1}, y)}{n}\end{aligned} \tag{69}$$

Since the loss function $L(., y)$ is ρ-Lipschitz, we have $L_x(h_1, y) - L_x(h_2, y) \leq \rho\|h_1 - h_2\|_1$ for any data point x and any two models h_1 and h_2. Plugging Eqn. 69 and the ρ-Lipschitz property back in Eqn. 68,

$$\begin{aligned}&\mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1'}(h^{min}_{a,S_1'}, y) \leq \frac{2}{n}\rho\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1\\ &\mathcal{L}_{S_1}(h^{min}_{a,S_1'}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y) \leq \frac{2}{n}\rho\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1\\ \Rightarrow &\max_{S,S'}\mathcal{L}_{S'}(h^{min}_{a,S}, y) - \mathcal{L}_S(h^{min}_{a,S}, y)\\ &\leq (\mathcal{L}_{S_1'}(h^{min}_{a,S_1}, y) - \mathcal{L}_{S_1'}(h^{min}_{a,S_1'}, y)) + (\mathcal{L}_{S_1}(h^{min}_{a,S_1'}, y) - \mathcal{L}_{S_1}(h^{min}_{a,S_1}, y))\\ &\leq \frac{2}{n}\rho\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1 + \frac{2}{n}\rho\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1\\ &= \frac{4}{n}\rho\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1\end{aligned} \tag{70}$$

Finally, combining with Eqn. 67, we obtain,

$$\begin{aligned}\max_{S,S'}\left\|h^{min}_{c,S'} - h^{min}_{c,S}\right\|_1^2 &\leq \frac{8\rho}{\lambda n}\left\|h^{min}_{a,S_1} - h^{min}_{a,S_1'}\right\|_1\\ &\leq \frac{8\rho}{\lambda n}\max_{S,S'}\left\|h^{min}_{a,S} - h^{min}_{a,S'}\right\|_1\\ &\leq \max_{S,S'}\left\|h^{min}_{a,S} - h^{min}_{a,S'}\right\|_1\end{aligned} \tag{71}$$

where the last inequality holds for a sufficiently large n such that $\frac{8\rho}{\lambda n} \leq 1$. If $\max_{S,S'}\left\|h^{min}_{c,S'} - h^{min}_{c,S}\right\|_1 \geq 1$, the result follows. Otherwise, we need a larger n such that $n \geq \frac{8\rho}{\lambda\max_{S,S'}\left\|h^{min}_{c,S'} - h^{min}_{c,S}\right\|_1}$. In both cases, we obtain,

$$\max_{S,S'}\left\|h^{min}_{c,S} - h^{min}_{c,S'}\right\|_1 \leq \max_{S,S'}\left\|h^{min}_{a,S} - h^{min}_{a,S'}\right\|_1 \tag{72}$$

Thus sensitivity of a causal learning function is lower than that of an associational learning function, that is, $\Delta\mathcal{F}_c \leq \Delta\mathcal{F}_a$. □

C Differential Privacy Guarantees with Tighter Data-dependent Bounds

In this section we present the differential privacy guarantee of a causal model based on a recent method (Papernot et al., 2017) that provides tighter data-dependent bounds.

Before analyzing the differential privacy guarantee, we provide a generalization result for a 0-1 classifier based on the results from Theorem 1: causal classification models trained on data from two different distributions $P(X)$ and $P^*(X)$ are more likely to output the same value for a new input than associational models.

Lemma 4. *Under the conditions of Theorem 1 and 0-1 loss, let $h^{min}_{c,S}$ be the loss-minimizing causal classification model trained on a dataset S from distribution P and let h^{min}_{c,S^*} be the loss-minimizing model trained on a dataset S^**

from $\mathtt{P}^*$. *Similarly, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}$ *be loss-minimizing associational classification models trained on* $\mathtt{S}$ *and* $\mathtt{S}^*$ *respectively. Then for any new data input* $\mathtt{x}$,

$$\min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big)$$
$$\geq \min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}(\mathtt{x})\big)$$

As the size of the training sample $|\mathtt{S}|=|\mathtt{S}^*|\to\infty$, *the LHS*$\to 1$.

Proof. Let $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,P}}=\arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{S}}(\mathtt{h},\mathtt{y})$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,P}^*}=\arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{S}^*}(\mathtt{h},\mathtt{y})$ be the loss-minimizing associational hypotheses under the two datasets $\mathtt{S}$ and $\mathtt{S}^*$ respectively, where $\mathcal{H}_{\mathtt{A}}$ is the set of hypotheses. We can analogously define $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,P}}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,P}^*}$. Likewise, let $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}=\arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{y})$ and similarly let $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}^*}=\arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y})$ be the loss-minimizing hypotheses over the two distributions. We can analogously define $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}$ and $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}$. For ease of exposition, let us consider a binary classification task where all associational models map $\mathtt{X}\to\{\mathtt{0},\mathtt{1}\}$ and causal models map $\mathtt{X}_{\mathtt{C}}\to\{\mathtt{0},\mathtt{1}\}$.

Infinite sample result. As $|\mathtt{S}|=|\mathtt{S}^*|\to\infty$, each of the models on $\mathtt{S}$ and $\mathtt{S}^*$ approach their loss-minimizing functions on the distributions $\mathtt{P}$ and $\mathtt{P}^*$ respectively. Then, for any input $\mathtt{x}$,

$$\lim_{|\mathtt{S}|\to\infty}\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}\qquad\lim_{|\mathtt{S}^*|\to\infty}\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}^*}\tag{73}$$
$$\lim_{|\mathtt{S}|\to\infty}\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}\qquad\lim_{|\mathtt{S}^*|\to\infty}\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}\tag{74}$$

From Theorem 1 (Equation 19), we know that $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}$. Therefore, for any new input $\mathtt{x}$ for a causal model, we obtain $\Pr\big(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}(\mathtt{x})=\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}(\mathtt{x})\big)=\mathtt{1}$, but not necessarily for associational models. This leads to,

$$\lim_{|\mathtt{S}|\to\infty,|\mathtt{S}^*|\to\infty}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big)=\mathtt{1}\tag{75}$$
$$\geq\lim_{|\mathtt{S}|\to\infty,|\mathtt{S}^*|\to\infty}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}(\mathtt{x})\big)\tag{76}$$

Finite sample result. Under finite samples, let $\mathtt{S}_1$ and $\mathtt{S}_2^*$ be the two datasets from $\mathtt{P}$ and $\mathtt{P}^*$ respectively that lead to the minimum probability of agreement between the two causal models $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}$.

$$\min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big)=\Pr\Big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}(\mathtt{x})\Big)\tag{77}$$

Now consider the probability of agreement for the two associational models trained on the same datasets, $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1^*}$. Without loss of generality, we can write the associational models as,

$$\begin{aligned}\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}(\mathtt{x})&=|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}(\mathtt{x})-\mathtt{h}_{\mathtt{a,S}_1}(\mathtt{x})|\\\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1^*}(\mathtt{x})&=|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}(\mathtt{x})-\mathtt{h}_{\mathtt{a,S}_1^*}(\mathtt{x})|\end{aligned}\tag{78}$$

where $\mathtt{h}_{\mathtt{a,S}_1}:\mathtt{X}\to\{\mathtt{0},\mathtt{1}\}$ and $\mathtt{h}_{\mathtt{a,S}_1^*}:\mathtt{X}\to\{\mathtt{0},\mathtt{1}\}$ are any two functions. Effectively, when $\mathtt{h}_{\mathtt{a,S}_1}$ is $\mathtt{1}$, it flips the output of the loss-minimizing associational model compared to the loss-minimizing causal model on $\mathtt{S}_1$ (and similarly for $\mathtt{h}_{\mathtt{a,S}_1^*}$ on $\mathtt{S}_1^*$).

Now we can select a different $\mathtt{S}_2^*\sim\mathtt{P}_2^*$ where $\mathtt{y}$ and the causal features remain the same as $\mathtt{S}_1^*$ but associational features are changed for each input $\mathtt{x}\in\mathtt{S}_2^*$. Therefore we construct $\mathtt{S}_2^*$ such that $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_2^*}$ but the term in the loss-minimizing associational model $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_2^*}$ has the following property: $\mathtt{h}_{\mathtt{a,S}_2^*}\neq\mathtt{h}_{\mathtt{a,S}_1}(\mathtt{x})$ if $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}$, and $\mathtt{h}_{\mathtt{a,S}_2^*}=\mathtt{h}_{\mathtt{a,S}_1}(\mathtt{x})$ if $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}\neq\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}$. Thus, under $\mathtt{S}_2^*$, for all $\mathtt{x}$,

$$\begin{aligned}&\Big|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_2^*}(\mathtt{x})\Big|\geq\Big|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_2^*}(\mathtt{x})\Big|\\&=\Big|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}_1^*}(\mathtt{x})\Big|=\max_{\mathtt{S},\mathtt{S}^*}\big|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big|\end{aligned}\tag{79}$$

Therefore, the disagreement between two associational models trained on two datasets is greater than or equal to the disagreement between causal models on the worst-case $\mathtt{S}_1$ and $\mathtt{S}_1^*$. Since the loss is 0-1 loss, the worst-case probability of agreement is lower.

$$\begin{aligned}&\max_{\mathtt{S},\mathtt{S}^*}\big|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big|\leq\max_{\mathtt{S},\mathtt{S}^*}\big|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}(\mathtt{x})-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}(\mathtt{x})\big|\\\Rightarrow&\min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}^*}(\mathtt{x})\big)\\&\geq\min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}^*}(\mathtt{x})\big)\end{aligned}$$

□

Based on the above generalization property, we now show that causal models provide stronger differential privacy guarantees than corresponding associational models. We utilize the *subsample and aggregate* technique (Dwork et al., 2014) that was extended for machine learning in Hamm et al. (2016) and Papernot et al. (2017), for constructing a differentiably private model. The framework considers $\mathtt{M}$ arbitrary teacher models that are trained on a separate subsample of the dataset without replacement. Then, a student model is trained on some auxiliary unlabeled data with the (pseudo) labels generated from a majority vote of the teachers. Differential privacy can be achieved by either perturbing the number of votes for each class (Papernot et al., 2017), or perturbing the learnt parameters of the student model (Hamm et al., 2016). For any new input, the output of the model is a majority vote on the predicted labels from the $\mathtt{M}$ models. The privacy guarantees are better if a larger number of teacher models agree on each input, since by definition the majority decision could not have been changed by modifying a single data point (or a single teacher's vote). Since causal models generalize to new distributions, intuitively we expect causal models trained on separate samples to agree more. Below we show that for a fixed amount of noise, a causal model is ϵ_c-DP compared to ϵ-DP for a associational model, where $\epsilon_c \leq \epsilon$.

Theorem 4. *Let* $\mathtt{D}$ *be any dataset generated from possibly a mixture of different distributions* $\Pr(\mathtt{X}, \mathtt{Y})$ *such that* $\Pr(\mathtt{Y}|\mathtt{X_C})$ *remains the same. Let* $\mathtt{n_k}$ *be the votes for the* $\mathtt{k}$th *class from* $\mathtt{M}$ *teacher models. Let* $\mathcal{M}$ *be the mechanism that produces a noisy max,* $\arg\max_{\mathtt{k}}\{\mathtt{n_k} + \mathtt{Lap}(2/\gamma)\}$*. Then the privacy budget* $\epsilon_{\mathtt{c}}$ *for a causal model trained on* $\mathtt{D}$ *is lower than that for an associational model.*

Proof. Consider a change in a single input example (x, y), leading to a new D' dataset. Since sub-datasets are sampled without replacement, only a single teacher model can change in D'. Let n'_k be the vote counts for a class k under D'. Because the change in a single input can only affect one model's vote, $|n_k - n'_k| \leq 1$.

Let the noise added to each class be $r_k \sim Lap(2/\gamma)$. Let the majority class (class with the highest votes) using data from D be i and the class with the second largest votes be j. Let us consider the minimum noise r^* required for class i to be the majority output under $\mathcal{M}$ over D. Then,

$$n_i + r^* > n_j + r_j$$

For i to have the maximum votes using $\mathcal{M}$ over D' too, we need,

$$n'_i + r_i > n'_j + r_j$$

In the worst case, $n'_i = n_i - 1$ and $n'_j = n_j + 1$ for some j. Thus, we need,

$$n_i - 1 + r_i > n_j + 1 + r_j \Rightarrow n_i + r_i > n_j + 2 + r_j \tag{80}$$

which shows that $r_i > r^* + 2$. Note that $r^* > r_j - (n_i - n_j)$. We have two cases:

CASE I: The noise $r_j < n_i - n_j$, and therefore $r^* < 0$. Writing $\Pr(i|D')$ to denote the probability that class i is chosen as the majority class under D',

$$\begin{aligned} \mathtt{P(i|D')} = \mathtt{P(r_i} \geq \mathtt{r^*} + 2) &= 1 - 0.5\exp(\gamma)\exp\left(\frac{1}{2}\gamma\mathtt{r^*}\right) \\ &= 1 - \exp(\gamma)(1 - \mathtt{P(r_i} \geq \mathtt{r^*})) \\ &= 1 - \exp(\gamma)(1 - \mathtt{P(i|D)}) \end{aligned} \tag{81}$$

where the equations on the right are due to Laplace c.d.f. Using the above equation, we can write:

$$\begin{aligned} \frac{\mathtt{P(i|D')}}{\mathtt{P(i|D)}} &= \exp(\gamma) + \frac{1 - \exp(\gamma)}{\mathtt{P(i|D)}} \\ &= \exp(\gamma) + \frac{1 - \exp(\gamma)}{\mathtt{P(r_i} \geq \mathtt{r^*})} \leq \exp(\epsilon) \end{aligned} \tag{82}$$

for some $\epsilon > 0$. As $P(i|D) = P(r_i \geq r^*)$ increases, the ratio decreases and thus the effective privacy budget (ϵ) decreases. Thus, a DP-mechanism based on teacher models with lower r^* (effectively higher $|r^*|$) will exhibit the lowest ϵ.

Below we show that the worst-case $|r^*|$ across any two datasets $\mathtt{D_1} \sim \mathtt{P}$, $\mathtt{D_2} \sim \mathtt{P^*}$ such that $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$ is higher for a causal model, and thus $\max_{\mathtt{D}} \mathtt{P(r_i} \geq \mathtt{r^*})$ is higher. Intuitively, $|r^*|$ is higher when there is more consensus between the $\mathtt{M}$ teacher models since $|r^*|$ is the difference between the votes for the highest voted class with the votes for the second-highest class. For two sub-datasets $\mathtt{D_1} \subset \mathtt{D}$ and $\mathtt{D_2} \subset \mathtt{D}$, let the two causal teacher models be $\mathtt{h_{c,D_1}}$ and $\mathtt{h_{c,D_2}}$, and the two associational teacher models be $\mathtt{h_{a,D_1}}$ and $\mathtt{h_{a,D_2}}$. From Lemma 4, for any new $\mathtt{x}$, there is more consensus among causal models. For any dataset $\mathtt{D}$,

$$\min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{c,D_1}(x)} = \mathtt{h_{c,D_2}(x)}) \geq \min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{a,D_1}(x)} = \mathtt{h_{a,D_2}(x)})$$

Hence worst-case $r_c^* \leq r^*$. From Equation 82, $\epsilon_c \leq \epsilon$.

CASE II: The noise $r_j >= n_i - n_j$, and therefore $r^* >= 0$. Following the steps above, we obtain:

$$\begin{aligned} \mathtt{P(i|D')} = \mathtt{P(r_i} \geq \mathtt{r^*} + 2) &= 0.5\exp(-\gamma)\exp\left(-\frac{1}{2}\gamma\mathtt{r^*}\right) \\ &= \exp(-\gamma)(\mathtt{P(r_i} \geq \mathtt{r^*})) \\ &= \exp(-\gamma)(\mathtt{P(i|D)}) \end{aligned} \tag{83}$$

Thus, the ratio does not depend on r^*.

$$\frac{\mathtt{P(i|D')}}{\mathtt{P(i|D)}} = \exp(-\gamma) \tag{84}$$

Under CASE II when the noise is higher to the differences in votes between the highest and second highest voted class, causal models provide the same privacy budget as associational models.

Hence, overall, $\epsilon_c \leq \epsilon$. □

D Maximum Advantage of a Differentially Private algorithm

Theorem 6. *Under the conditions of Theorem 1, let* $\mathtt{S} \sim \mathtt{P(X, Y)}$ *be a dataset sampled from* $\mathtt{P}$*. Let* $\hat{\mathcal{F}}_{\mathtt{c,S}}$ *and* $\hat{\mathcal{F}}_{\mathtt{a,S}}$ *be the differentially private mechanisms trained on* $\mathtt{S}$ *by adding identical Laplacian noise to the causal and associational learning functions from Lemma 1 respectively. Assume that a membership inference adversary is provided inputs sampled from either* $\mathtt{P}$ *or* $\mathtt{P^*}$*, where* $\mathtt{P^*}$ *is any distribution such that* $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$*. Then, across all adversaries* $\mathcal{A}$

that predict membership in $\mathtt{S} \sim \mathtt{P}$*, the worst-case membership advantage of* $\hat{\mathcal{F}}_{\mathtt{c},\mathtt{S}}$ *is not greater than that of* $\hat{\mathcal{F}}_{\mathtt{a},\mathtt{S}}$.

$$\max_{\mathcal{A},\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}_{\mathtt{c},\mathtt{S}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) \le \max_{\mathcal{A},\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}_{\mathtt{a},\mathtt{S}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*)$$

Proof. Consider a neighboring dataset $\mathtt{S}'$ to $\mathtt{S} \sim \mathtt{P}$ such that $\mathtt{S}'$ replaces data point $\mathtt{x} \in \mathtt{S}$ with a different point $\mathtt{x}'$. Following Theorem 1 proof from Yeom et al. (2018), the membership advantage of an adversary $\mathcal{A}$ on a differentially private algorithm $\hat{\mathcal{F}}$ can be written as:

$$\begin{aligned}\mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) &= \Pr(\mathcal{A} = \mathtt{1} | \mathtt{b} = \mathtt{1}) - \Pr(\mathcal{A} = \mathtt{1} | \mathtt{b} = \mathtt{0}) \\ &= \Pr\big(\mathcal{A}(\mathtt{x}, \hat{\mathcal{F}}_{\mathtt{S}}) = \mathtt{1} | \mathtt{x} \in \mathtt{S}\big) - \Pr\big(\mathcal{A}(\mathtt{x}, \hat{\mathcal{F}}_{\mathtt{S}'}) = \mathtt{1} | \mathtt{x} \in \mathtt{S}\big)\end{aligned} \tag{85}$$

where $\mathcal{A}(\mathtt{x}, \hat{\mathcal{F}}_{\mathtt{S}})$ denotes a membership adversary for an algorithm $\hat{\mathcal{F}}_{\mathtt{S}}$ trained on a dataset $\mathtt{S}$, and $\mathcal{A}(\mathtt{x}, \hat{\mathcal{F}}_{\mathtt{S}'})$ denotes an adversary for algorithm $\hat{\mathcal{F}}_{\mathtt{S}'}$ trained on $\mathtt{S}'$. Without loss of generality for the case where there are an infinite number of models $\mathtt{h}$, assume that models are sampled from a discrete set of K models: $\{\mathtt{h}_1, \mathtt{h}_2, ..., \mathtt{h}_{\mathtt{K}}\}$. Then using the law of total probability over the models yielded by the algorithms $\hat{\mathcal{F}}_{\mathtt{S}}$ and $\hat{\mathcal{F}}_{\mathtt{S}'}$,

$$\begin{aligned}\mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) &= \sum_{\mathtt{j}=\mathtt{1}}^{\mathtt{K}} \Pr\big(\mathcal{A}(\mathtt{x}, \mathtt{h}_{\mathtt{j}}) = \mathtt{1}\big) \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) \\ &- \sum_{\mathtt{j}=\mathtt{1}}^{\mathtt{K}} \Pr\big(\mathcal{A}(\mathtt{x}, \mathtt{h}_{\mathtt{j}}) = \mathtt{1}\big) \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big) \\ &= \sum_{\mathtt{j}=\mathtt{1}}^{\mathtt{K}} \Pr\big(\mathcal{A}(\mathtt{x}, \mathtt{h}_{\mathtt{j}}) = \mathtt{1}\big) [\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big)]\end{aligned} \tag{86}$$

where $\Pr\big(\mathcal{A}(\mathtt{x}, \mathtt{h}_{\mathtt{j}}) = \mathtt{1}\big)$ can be interpreted as the non-negative weights in a sum. Thus, the above is a weighted sum and will be maximum when positive values for $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big)$ have the highest weight and negative values for $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big)$ have zero weight. It follows that to obtain the maximum advantage, the adversary will choose $\Pr\big(\mathcal{A}(\mathtt{x}, \mathtt{h}_{\mathtt{j}}) = \mathtt{1}\big) = \mathtt{1}$ if $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big) > \mathtt{0}$, and 0 otherwise. In other words, the adversary predicts membership in train set for an input $\mathtt{x} \in \mathtt{S}$ whenever probability of the given model $\mathtt{h}_{\mathtt{j}}$ being generated from $\hat{\mathcal{F}}_{\mathtt{S}}$ is higher than it being generated from $\hat{\mathcal{F}}_{\mathtt{S}'}$.

Let $\mathtt{H}_+ \subset \mathtt{H}$ be the set of models for which $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big) > \mathtt{0}$. Similarly, let $\mathtt{H}_- = \mathtt{H} \setminus \mathtt{H}_+$ be the set of models that are more probable to be generated from $\hat{\mathcal{F}}_{\mathtt{S}'}$: $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) \ge \mathtt{0}$. The worst-case adversary selects datasets $\mathtt{S} \sim \mathtt{P}, \mathtt{S}'$ such that the sum $\sum_{\mathtt{h}_{\mathtt{j}} \in \mathtt{H}_+} \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}\big) - \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} = \mathtt{h}_{\mathtt{j}}\big)$ is the highest. Therefore, for a given distribution $\mathtt{P}$ and a differentially private algorithm $\hat{\mathcal{F}}_{\mathtt{S}}$ learnt on $\mathtt{S} \sim \mathtt{P}$, we can write the maximum membership advantage as,

$$\begin{aligned}&\max_{\mathcal{A},\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}_{\mathtt{S}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) \\ &= \mathtt{max}_{\mathtt{S},\mathtt{S}'} \sum_{\mathtt{h}_{\mathtt{j}} \in \mathtt{H}_+} [\mathtt{P}(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}) - \mathtt{P}(\hat{\mathcal{F}}'_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}})] \\ &= \mathtt{max}_{\mathtt{S},\mathtt{S}'} \mathtt{P}(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+) - \mathtt{P}(\hat{\mathcal{F}}'_{\mathtt{S}} \in \mathtt{H}_+) \\ &= \mathtt{max}_{\mathtt{S},\mathtt{S}'} \mathtt{P}(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+) - (\mathtt{1} - \mathtt{P}(\hat{\mathcal{F}}'_{\mathtt{S}} \in \mathtt{H}_-)) \\ &= \mathtt{max}_{\mathtt{S},\mathtt{S}'} 2 \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+\big) - \mathtt{1}\end{aligned} \tag{87}$$

where the last equality is since $\hat{\mathcal{F}}_{\mathtt{S}}$ and $\hat{\mathcal{F}}_{\mathtt{S}'}$ have Laplace noise added from identical distributions and thus $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+\big) = \Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'} \in \mathtt{H}_-\big)$. Equation 87 provides the maximum membership advantage for any ϵ-DP mechanism $\hat{\mathcal{F}}_{\mathtt{S}}$ with Laplace noise.

We next show that Eqn. 87 for a causal differentially private mechanism $\hat{\mathcal{F}}_{\mathtt{c}}$ is not greater than that for an associational mechanism $\hat{\mathcal{F}}_{\mathtt{a}}$. Let $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}}\big)$ be a Laplace distribution with mean at $\mathtt{h}_{\mathtt{S},\mathtt{min}}$ and $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}'}\big)$ be a Laplace distribution with mean $\mathtt{h}_{\mathtt{S}',\mathtt{min}}$ with identical scale/noise parameters. We would like to find the boundary model $\mathtt{h}^\dagger$ of the set $\mathtt{H}_+$ where $\mathtt{P}(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}}) = \mathtt{P}(\hat{\mathcal{F}}'_{\mathtt{S}} = \mathtt{h}_{\mathtt{j}})$, since $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+\big)$ is the probability under the Laplace distribution, cut off at the point $\mathtt{h}^\dagger$. Due to identical noise for $\hat{\mathcal{F}}_{\mathtt{S}}$ and $\hat{\mathcal{F}}'_{\mathtt{S}}$ and the symmetry of the Laplace distribution, the boundary $\mathtt{h}^\dagger$ corresponds to the midpoint of $\mathtt{h}_{\mathtt{S},\mathtt{min}}$ and $\mathtt{h}_{\mathtt{S}',\mathtt{min}}$: $\mathtt{0.5}(\mathtt{h}_{\mathtt{S},\mathtt{min}} + \mathtt{h}_{\mathtt{S}',\mathtt{min}})$. Alternatively, the ℓ_1 distance of the boundary $\mathtt{h}^\dagger$ from the means of the Laplace distributions can be written as (for a worst case $\mathtt{S}, \mathtt{S}'$),

$$\left\|\mathtt{h}^\dagger - \mathtt{h}_{\mathtt{S},\mathtt{min}}\right\|_1 = \frac{\|\mathtt{h}_{\mathtt{S}',\mathtt{min}} - \mathtt{h}_{\mathtt{S},\mathtt{min}}\|_1}{2} = \frac{\Delta\mathcal{F}_{\mathtt{S}}}{2} \tag{88}$$

where $\Delta\mathcal{F}_{\mathtt{S}}$ is the sensitivity of the learning function $\mathcal{F}_{\mathtt{S}}$ and the last equality is due to the choice of worst-case $\mathtt{S}$ and $\mathtt{S}'$.

From Lemma 1, we know that sensitivity of a causal learning function is lower than that of an associational learning function.

$$\Delta\mathcal{F}_{\mathtt{c},\mathtt{S}} \le \Delta\mathcal{F}_{\mathtt{a},\mathtt{S}} \tag{89}$$

Thus, ℓ_1 distance of $\mathtt{h}^\dagger_{\mathtt{c}}$ from the mean $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ is lower for a causal learning function, and consequently its probability $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} = \mathtt{h}^\dagger_{\mathtt{c}}\big)$ is higher. Now the set $\mathtt{H}_+$ is a one-sided boundary on the values of $\mathtt{h}$ and includes the mean of the Laplace distribution. Given symmetry of the Laplace distribution, probability of $\hat{\mathcal{F}}_{\mathtt{S}}$ lying in H_+, $\Pr\big(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+\big)$ should be lower whenever the probability at the one-sided boundary is higher. Therefore, $\mathtt{P}(\hat{\mathcal{F}}_{\mathtt{S}} \in \mathtt{H}_+)$ is lower for a causal mechanism than the associational learning mechanism.

$$\Delta\mathcal{F}_{\mathtt{c},\mathtt{S}} \le \Delta\mathcal{F}_{\mathtt{a},\mathtt{S}} \Rightarrow \Pr\big(\hat{\mathcal{F}}_{\mathtt{c},\mathtt{S}} \in \mathtt{H}_+\big) \le \Pr\big(\hat{\mathcal{F}}_{\mathtt{a},\mathtt{S}} \in \mathtt{H}_+\big) \tag{90}$$

Finally, using the above equation in Eqn. 87 shows that the maximum membership advantage of a causal model is lower.

$$\max_{\mathcal{A},\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}_{\mathtt{c},\mathtt{S}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) \leq \max_{\mathcal{A},\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}, \hat{\mathcal{F}}_{\mathtt{a},\mathtt{S}}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) \tag{91}$$

□

E Infinite Sample Robustness to MI Attacks

Corollary 2. *Under the conditions of Theorem 1, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ *be a causal model trained using empirical risk minimization on a dataset* $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$ *with sample size* $\mathtt{n}$*. As* $\mathtt{n} \to \infty$*, membership advantage* $\mathtt{Adv}(\mathcal{A}, \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}) \to 0$.

Proof. $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ can be obtained by empirical risk minimization.

$$\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{C}}} \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h}, \mathtt{y}) = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{C}}} \frac{1}{\mathtt{n}} \sum_{\mathtt{i}=1}^{\mathtt{n}} \mathtt{L}_{\mathtt{x}_\mathtt{i}}(\mathtt{h}, \mathtt{y}) \tag{92}$$

As $|\mathtt{S}| = \mathtt{n} \to \infty$, $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} \to \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}$. Suppose now that there exists another $\mathtt{S}'$ of the same size such that $\mathtt{S}' \sim \mathtt{P}^*$. Then as $|\mathtt{S}'| \to \infty$, $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'} \to \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$.

From Theorem 1, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$. Thus,

$$\lim_{\mathtt{n}\to\infty} \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} = \lim_{\mathtt{n}\to\infty} \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'} \tag{93}$$

Equation 93 implies that as $\mathtt{n} \to \infty$, the learnt $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ does not depend on the training set, as long as the training set is sampled from any distribution $\mathtt{P}^*$ such that $\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{C})$. That is, being the global minimizer over distributions, $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}$ does not depend on its training set. Therefore, $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}(\mathtt{x})$ is independent of whether $\mathtt{x}$ is in the training set.

$$\begin{aligned}
\lim_{\mathtt{n}\to\infty} \mathtt{Adv}(\mathcal{A}, \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}) &= \Pr(\mathcal{A} = 1|\mathtt{b} = 1) - \Pr(\mathcal{A} = 1|\mathtt{b} = 0) \\
&= \mathbb{E}[\mathcal{A}|\mathtt{b} = 1] - \mathbb{E}[\mathcal{A}|\mathtt{b} = 0] \\
&= \mathrm{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}})|\mathtt{b} = 1] - \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}})|\mathtt{b} = 0] \\
&= \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}})] - \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}})] = 0
\end{aligned} \tag{94}$$

where the second last equality follows since any function of $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ is independent of the training dataset. □

F Robustness to Attribute Inference Attacks

Theorem 7. *Given a dataset* $\mathtt{S}(\mathtt{X}, \mathtt{Y})$ *of size* $\mathtt{n}$ *and a structural causal model that connects* $\mathtt{X}$ *to* $\mathtt{Y}$*, a causal model* $\mathtt{h}_\mathtt{c}$ *makes it impossible to infer non-causal features.*

Proof. The proof follows trivially from the definition of a causal model. $\mathtt{h}_\mathtt{c}$ includes only causal features during training. Thus, $\mathtt{h}(\mathtt{x})$ is independent of all features not in $\mathtt{X}_\mathtt{c}$.

$$\begin{aligned}
\mathtt{Adv}(\mathcal{A}, \mathtt{h}) &= \mathtt{Pr}(\mathcal{A} = 1|\mathtt{x}_\mathtt{s} = 1) - \mathtt{Pr}(\mathcal{A} = 1|\mathtt{x}_\mathtt{s} = 0) \\
&= \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1|\mathtt{x}_\mathtt{s} = 1) - \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1|\mathtt{x}_\mathtt{s} = 0) \\
&= \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1) - \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1) = 0
\end{aligned}$$

□

G Generating Train and Test Datasets

This section describes how the train and test data are generated for the target and attacker models. The target model is built using a dataset generated using the bnlearn library. We divide the total dataset into training and test datasets in a 60 : 40 ratio.

To generate train and test sets for the attacker model, the output of the target model for each of the training and test dataset is again divided into 50:50 ratio. Note that the attacker model is trained on the confidence outputs of the target model. Therefore, the training set for the attacker model consists of output confidence values from the target model's training as well as the test dataset. The dataset creation process is summarized in Figure 5.

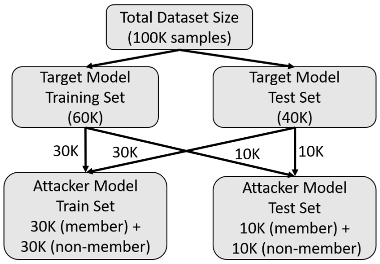

Figure 5: Dataset division for training target and attacker models.