Alleviating Privacy Attacks via Causal Learning

Shruti Tople [1] **Amit Sharma** [1] **Aditya V. Nori** [1]

Abstract

Machine learning models, especially deep neural networks have been shown to be susceptible to privacy attacks such as *membership inference* where an adversary can detect whether a data point was used for training a black-box model. Such privacy risks are exacerbated when a model's predictions are used on an unseen data distribution. To alleviate privacy attacks, we demonstrate the benefit of predictive models that are based on the causal relationships between input features and the outcome. We first show that models learnt using causal structure generalize better to unseen data, especially on data from different distributions than the train distribution. Based on this generalization property, we establish a theoretical link between causality and privacy: compared to associational models, causal models provide stronger differential privacy guarantees and are more robust to membership inference attacks. Experiments on simulated Bayesian networks and the colored-MNIST dataset show that associational models exhibit upto 80% attack accuracy under different test distributions and sample sizes whereas causal models exhibit attack accuracy close to a random guess.

1 Introduction

Machine learning algorithms, especially deep neural networks (DNNs) have found diverse applications in various fields such as healthcare (Esteva et al., 2019), gaming (Mnih et al., 2013), and finance (Tsantekidis et al., 2017; Fischer & Krauss, 2018). However, a line of recent research has shown that deep learning algorithms are susceptible to privacy attacks that leak information about the training dataset (Fredrikson et al., 2015; Rahman et al., 2018; Song & Shmatikov, 2018; Hayes et al., 2017). Particularly, one such attack called *membership inference* reveals whether a data sample was present in the training dataset (Shokri et al., 2017). The privacy risks due to membership inference elevate when the DNNs are trained on sensitive data such as in healthcare applications. For example, HIV patients would not want to reveal their participation in the training dataset.

Membership inference attacks are shown to exploit overfitting of the model on the training dataset (Yeom et al., 2018). Existing defenses propose the use of generalization techniques such as adding learning rate decay, dropout or using adversarial regularization techniques (Nasr et al., 2018b; Salem et al., 2018). All these approaches assume that the test and the training data belong to the same distribution. In practice, a model trained using data from one distribution is often used on a (slightly) different distribution. For example, hospitals in one region may train a model and share it with hospitals in different regions. However, generalizing to a new context is a challenge for any machine learning model. We extend the scope of membership inference attacks to different distributions and show that the risk increases for associational models as the test distribution is changed.

Our Approach. To alleviate privacy attacks, we propose using models that depend on the causal relationship between input features and the output. Causal learning has been used to optimize for fairness and explainability properties of the predicted output (Kusner et al., 2017; Nabi & Shpitser, 2018; Datta et al., 2016). However, the connection of causal learning to enhancing privacy of models is yet unexplored. To the best of our knowledge, we provide the first analysis of privacy benefits of causal models. By definition, causal relationships are invariant across input distributions (Peters et al., 2016), and therefore predictions of *causal models* should be independent of the observed data distribution, let alone the observed dataset. Thus, causal models generalize better even with changes in the data distribution.

In this paper, we show that the generalizability property of causal models directly ensures better privacy guarantees for the input data. Concretely, we prove that with reasonable assumptions, **a causal model always provides stronger (i.e., smaller ϵ value) differential privacy guarantees than an associational model trained on the same features and with the same amount of added noise**. Consequently, we

[1] Microsoft Research. Correspondence to: Shruti Tople <shruti.tople@microsoft.com>, Amit Sharma <amshar@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

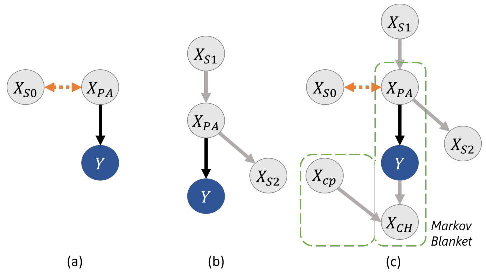


Figure 1: Structural causal model where a directed edge denotes a causal relationship; dashed bidirectional edges denote correlation. A causal predictive model includes only the parents of $\mathtt{Y}$: $\mathtt{X}_{\mathtt{PA}}$ [(a) and (b)]. Panel (c) shows the Markov Blanket of $\mathtt{Y}$.

show that membership attacks are ineffective (almost a random guess) on causal models trained on infinite samples.

Empirical attack accuracies on four different tabular datasets and the colored MNIST image dataset (Arjovsky et al., 2019) confirm our theoretical claims. On tabular data, we find that 60K training samples are sufficient to reduce the attack accuracy of a causal model to a random guess. In contrast, membership attack accuracy for neural network-based associational models increases up to 80% as the test distribution is changed. On colored MNIST dataset, we find that attack accuracy for causal model is close to a random guess (50%) compared to 66% for an associational model under a shift in the data distribution.

To summarize, our main contributions include:

- For the same amount of added noise, models learned using causal structure provide stronger ϵ-differential privacy guarantees than corresponding associational models.
- Models trained using causal features are *provably* more robust to membership inference attacks than typical associational models such as neural networks.
- On the colored MNIST dataset and simulated Bayesian Network datasets where the test distribution may not be the same as the training distribution, the membership inference attack accuracy of causal models is close to a "random guess" (i.e., 50%) whereas associational models exhibit 65-80% attack accuracy.

2 Generalization Property of Causal Models

Causal models generalize well since their output depends on stable, causal relationships between input features and the outcome instead of associations between them (Peters et al., 2016). Our goal is to study the effect of this generalization property on privacy of training data.

2.1 Background: Causal Model

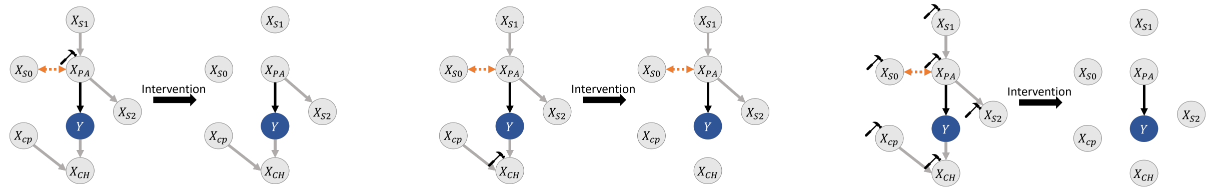
Intuitively, a causal model identifies a subset of features that have a causal relationship with the outcome and learns a function from the subset to the outcome. To construct a causal model, one may use a structural causal graph based on domain knowledge that defines causal features as parents of the outcome under the graph. Alternatively, one may exploit the strong relevance property from (Pellet & Elisseeff, 2008), use score-based learning algorithms (Scutari, 2009) or recent methods for learning invariant relationships from training datasets from different distributions (Peters et al., 2016; Arjovsky et al., 2019; Bengio et al., 2019; Mahajan et al., 2020), or learn based on a combination of randomized experiments and observed data. Note that this is different from training probabilistic graphical models, wherein an edge conveys an associational relationship. Further details on causal models are in (Pearl, 2009; Peters et al., 2017).

For ease of exposition, we assume the structural causal graph framework throughout. Consider data from a distribution $(\mathtt{X}, \mathtt{Y}) \sim \mathtt{P}$ where $\mathtt{X}$ is a k-dimensional vector. Our goal is to learn a function $\mathtt{h}(\mathtt{X})$ that predicts $\mathtt{Y}$. Figure 1 shows causal graphs that denote the different relationships between $\mathtt{X}$ and $\mathtt{Y}$. Nodes of the graph represent variables and a directed edge represents a direct causal relationship from a source to target node. Denote $\mathtt{X}_{\mathtt{PA}} \subseteq \mathtt{X}$, the parents of $\mathtt{Y}$ in the causal graph. Fig. (1a) shows the scenario where $\mathtt{X}$ contains variables $\mathtt{X}_{\mathtt{S0}}$ that are correlated to $\mathtt{X}_{\mathtt{PA}}$ in $\mathtt{P}$, but not necessarily connected to either $\mathtt{X}_{\mathtt{PA}}$ or $\mathtt{Y}$. These correlations may change in the future, therefore a generalizable model should not include these features. Similarly, Fig. (1b) shows parents and children of $\mathtt{X}_{\mathtt{PA}}$. The d-separation principle states that a node is independent of its ancestors conditioned on all its parents (Pearl, 2009). Thus, $\mathtt{Y}$ is independent of $\mathtt{X}_{\mathtt{S1}}$ and $\mathtt{X}_{\mathtt{S2}}$ conditional on $\mathtt{X}_{\mathtt{PA}}$. Including them in a model does not add predictive value (and further, avoids prediction error when the relationships between $\mathtt{X}_{\mathtt{S1}}$, $\mathtt{X}_{\mathtt{S2}}$ and $\mathtt{X}_{\mathtt{PA}}$ change).

The key insight is that building a model for predicting $\mathtt{Y}$ using its parents $\mathtt{X}_{\mathtt{PA}}$ ensures that the model generalizes to other distributions of $\mathtt{X}$, and also to changes in other causal relationships between $\mathtt{X}$, as long as the causal relationship of $\mathtt{X}_{\mathtt{PA}}$ to $\mathtt{Y}$ is stable. We call such a model a *causal* model, the features in $(\mathtt{X}_{\mathtt{C}} = \mathtt{X}_{\mathtt{PA}})$ the *causal features*, and assume that all causal features for $\mathtt{Y}$ are observed. In contrast, an *associational* model uses all the available features.

Here we would like to distinguish causal features from $\mathtt{Y}$'s Markov Blanket. The Markov Blanket (Pellet & Elisseeff, 2008) for $\mathtt{Y}$ contains its parents, children and parents of children. Conditioned on its Markov blanket (Fig. 1c), $\mathtt{Y}$ is independent of all other variables in the causal graph, and therefore past work (Aliferis et al., 2010) suggests to build a predictive model using the features in $\mathtt{Y}$'s Markov Blanket[1]. However, such a model is not robust to interventions. For instance, if there is an intervention on $\mathtt{Y}$'s children in a new

[1] In some cases, it may be necessary to use Y's children for prediction, e.g., in predicting disease based on its symptoms. However, such a model will not generalize under intervention— it makes an implicit assumption that symptoms will never be intervened upon, and that all causes of symptoms are observed.



(a) (b) (c)

Figure 2: Interventions on (a) parents of Y, (b) children of Y, and (c) all features. The black hammer denotes an intervention and each right subfigure shows the resultant causal model. Relationship between causal features and Y, $Y = f(X_{PA})$ remains invariant under all interventions but the relationship between other features and Y varies based on the intervention.

domain (Fig. 2b), it will break the correlation between Y and X_{CH} and lead to incorrect predictions. To summarize, Fig. (2c) demonstrates how a causal model based on parents is robust to all interventions on X, unlike an associational model built using the Markov Blanket or other features.

2.2 Generalization to New Distributions

We state the generalization property of causal models and show how it results in a stronger differential privacy guarantee. We first define *In-distribution* and *Out-of-distribution* generalization error. Throughout, $L(.,.)$ refers to the loss on a single input and $\mathcal{L}_P(.,.) = \mathbb{E}_P L(.,.)$ refers to the expected value of the loss over a distribution $P(X, Y)$. We refer to $h : X \rightarrow Y$ as the hypothesis function or simply the model. Then, $L(h, h')$ is a loss function quantifying the difference between any two models h and h'.

Definition 1. In-Distribution Generalization Error ($\mathtt{IDE}$). *Consider a dataset $S \sim P(X, Y)$. Then for a model $h : X \rightarrow Y$ trained on S, the in-distribution generalization error is given by:*

$$\mathtt{IDE}_P(h, y) = \mathcal{L}_P(h, y) - \mathcal{L}_{S\sim P}(h, y) \quad (1)$$

Definition 2. Out-of-Distribution Generalization Error ($\mathtt{ODE}$)**.** *Consider a dataset S sampled from a distribution $P(X, Y)$. Then for a model $h : X \rightarrow Y$ trained on S, the out-of-distribution generalization error with respect to another distribution $P^*(X, Y)$ is given by:*

$$\mathtt{ODE}_{P,P^*}(h, y) = \mathcal{L}_{P^*}(h, y) - \mathcal{L}_{S\sim P}(h, y) \quad (2)$$

Definition 3. Discrepancy Distance ($\mathtt{disc}_L$) **(Def. 4 in (Mansour et al., 2009))**. *Let $\mathcal{H}$ be a set of models, $h : X \rightarrow Y$. Let $L : Y \times Y \rightarrow \mathbb{R}_+$ define a loss function over Y for any such model h. Then the discrepancy distance $\mathtt{disc}_L$ over any two distributions $P(X, Y)$ and $P^*(X, Y)$ is given by:*

$$\mathtt{disc}_{L,\mathcal{H}}(P, P^*) = \max_{h,h'\in\mathcal{H}} |\mathcal{L}_P(h, h') - \mathcal{L}_{P^*}(h, h')| \quad (3)$$

Intuitively, the term $\mathtt{disc}_L(P, P^*)$ denotes the distance between the two distributions. Higher the distance, higher is the chance of an error when transferring model h from one distribution to another. Next we state the theorem on the generalization property of causal models.

Theorem 1. *Consider a structural causal graph G that connects X to Y, and causal features $X_C \subset X$ where X_C represent the parents of Y under G. Let $P(X, Y)$ and $P^*(X, Y)$ be two distributions with arbitrary $P(X)$ and $P^*(X)$, having overlap, $P(X = x) > 0$ whenever $P^*(X = x) > 0$. In addition, the causal relationship between X_C and Y is preserved, which implies that $P(Y|X_C) = P^*(Y|X_C)$. Let L be a symmetric loss function that obeys the triangle inequality (such as L1, L2 or 0-1 loss), and let $f : X_C \rightarrow Y$ be the optimal predictor among all hypotheses using X_C features under L, i.e., $f = \arg\min_h L_{x_c}(y, h(x_c))$ for all x_c, and thus f depends only on $\Pr(Y|X_C)$ (e.g., $f := \mathbb{E}[Y|X_C]$ for L2 loss). Further, assume that $\mathcal{H}_C$ represents the set of causal models $h_c : X_C \rightarrow Y$ that may use all causal features and $\mathcal{H}_A$ represent the set of associational models $h_a : X \rightarrow Y$ that may use all available features, such that $f \in \mathcal{H}_C$ and $\mathcal{H}_C \subseteq \mathcal{H}_A$.*

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

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

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

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

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

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

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

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

Proof Sketch. As an example, consider a colored MNIST data distribution $\mathtt{P}$ such that the label $\mathtt{Y}$ is assigned based on the shape of a digit. Here the shape features represent the causal features ($\mathtt{X_C}$). If the shape is closest to shapes for $\{0, 1, 2, 3, 4\}$ then $\mathtt{Y} = 0$, else $\mathtt{Y} = 1$. Additionally, all images classified as 1 are colored with the same color (say red). Then, under a suitably expressive class of models, the loss-minimizing associational model may use only the color feature to obtain zero error, while the loss-minimizing causal model still uses the shape (causal) features. On any new $\mathtt{P}^*$ that does not follow the same correlation of digits with color, we expect that the associational model will have higher error than the causal model.

Formally, since $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$ and $\mathtt{f} \in \mathcal{H}_\mathtt{C}$, the optimal causal model that minimizes loss over $\mathtt{P}$ is the same as the loss-minimizing model over $\mathtt{P}^*$. That is, $\mathtt{h^{OPT}_{c,P}} = \mathtt{h^{OPT}_{c,P^*}}$. However for associational models, the optimal models may not be the same $\mathtt{h^{OPT}_{a,P}} \neq \mathtt{h^{OPT}_{a,P^*}}$ and thus there is an additional loss term when generalizing to data from $\mathtt{P}^*$. The rest of the proof follows from triangle inequalities on the loss function and the standard bounds for $\mathtt{IDE}$ (in Suppl. Section A.1).

For individual instances, we present a similar result on the worst-case generalization error (proof in Suppl. Section A.2).

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

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

3 Main Result: Privacy with Causality

We now present our main result on the privacy guarantees and attack robustness of causal models.

3.1 Differential Privacy Guarantees

Differential privacy (Dwork et al., 2014) provides one of the strongest notions of privacy to hide the participation of an individual sample in the dataset. To state informally, it ensures that the presence or absence of a single data point in the input dataset does not change the output by much.

Definition 4 (Differential Privacy)**.** *A mechanism* $\mathtt{M}$ *with domain* $\mathcal{I}$ *and range* $\mathcal{O}$ *satisfies* ϵ*-differential privacy if for any two datasets* $d, d' \in \mathcal{I}$ *that differ only in one input and for a set* $\mathcal{S} \subseteq \mathcal{O}$*, the following holds:* $\Pr(\mathcal{M}(d) \in \mathcal{S}) \leq e^\epsilon \Pr(\mathcal{M}(d') \in \mathcal{S})$

The standard approach to designing a differentially private mechanism is by calculating the *sensitivity* of an algorithm and adding noise proportional to the sensitivity. Sensitivity captures the change in the output of a function due to changing a single data point in the input. Higher the sensitivity, larger is the amount of noise required to make any function differentially private with reasonable ϵ guarantees. Below we provide a formal definition of sensitivity, derive a corollary based on the generalization property from Theorem 2, and then show that sensitivity of a causal learning function is lower than or equal to an associational learning function (proofs are in Suppl. Section B).

Definition 5 (Sensitivity (From Def. 3.1 in (Dwork et al., 2014))**.** *Let* $\mathcal{F}$ *be a function that maps a dataset to a vector in* $\mathbb{R}^d$*. Let* $\mathtt{S}$, $\mathtt{S'}$ *be two datasets such that* $\mathtt{S'}$ *differs from* $\mathtt{S}$ *in one data point. Then the* l_1*-sensitivity of a function* $\mathcal{F}$ *is defined as:* $\Delta\mathcal{F} = \max_{\mathtt{S,S'}} ||\mathcal{F}(\mathtt{S}) - \mathcal{F}(\mathtt{S'})||_1$

Corollary 1. *Let* $\mathtt{S}$ *be a dataset of* $\mathtt{n}$ $\mathtt{(x, y)}$ *values, such that* $\mathtt{y^{(i)}} \sim \mathtt{P(Y|X_C = x^{(i)})} \forall \mathtt{(x^{(i)}, y^{(i)})} \in \mathtt{S}$*, where* $\mathtt{P(Y|X_C)}$ *is the invariant conditional distribution on the causal features* $\mathtt{X_C}$*. Consider a neighboring dataset* $\mathtt{S'}$ *such that* $\mathtt{S'} = \mathtt{S} \backslash \mathtt{(x, y)} + \mathtt{(x', y')}$ *where* $\mathtt{(x, y)} \in \mathtt{S}$*,* $\mathtt{(x', y')} \notin \mathtt{S}$*, and* $\mathtt{(x', y')}$ *shares the same conditional distribution* $\mathtt{y'} \sim \mathtt{P(Y|X_C = x'_c)}$*. 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.*

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

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* $\theta \in \Omega \subseteq \mathbb{R}^n$*. Let* $\mathtt{h^{min}_S}(\mathtt{x}; \theta_\mathtt{S})$ *be a model learnt using* $\mathtt{S}$ *as training data and* $\mathtt{h^{min}_{S'}}(\mathtt{x}; \theta_\mathtt{S'})$ *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* $\mathtt{h^{min}_{c,S}} \leftarrow \mathcal{F}_\mathtt{c}(\mathtt{S})$ *and* $\mathtt{h^{min}_{c,S'}} \leftarrow \mathcal{F}_\mathtt{c}(\mathtt{S'})$ *is lower than or equal to the sensitivity of an associational learning function* $\mathcal{F}_a$ *that outputs* $\mathtt{h^{min}_{a,S}} \leftarrow \mathcal{F}_\mathtt{a}(\mathtt{S})$ *and* $\mathtt{h^{min}_{a,S'}} \leftarrow \mathcal{F}_\mathtt{a}(\mathtt{S'})$*,*

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

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

We now prove our main result on differential privacy.

Theorem 3. *Let* $\hat{\mathcal{F}}_c$ *and* $\hat{\mathcal{F}}_a$ *be the differentially private mechanisms, obtained by adding Laplace noise to model parameters of the causal learning and associational learning functions* $\mathcal{F}_c$ *and* $\mathcal{F}_a$ *respectively. Let* $\hat{\mathcal{F}}_c$ *and* $\hat{\mathcal{F}}_a$ *provide* $\epsilon_\mathtt{c}$*-DP and* $\epsilon_\mathtt{a}$*-DP guarantees respectively. Then, for equivalent noise added to both the functions and sampled from the same distribution,* $\mathtt{Lap(Z)}$*, we have* $\epsilon_\mathtt{c} \leq \epsilon_\mathtt{a}$*.*

Proof. According to the Def. 3.3 of Laplace mechanism from (Dwork et al., 2014), we have,

$$\hat{\mathcal{F}}_c = \mathcal{F}_\mathtt{c} + \mathcal{K} \sim \mathtt{Lap}(\frac{\Delta\mathcal{F}_\mathtt{c}}{\epsilon_\mathtt{c}}) \qquad \hat{\mathcal{F}}_a = \mathcal{F}_\mathtt{a} + \mathcal{K} \sim \mathtt{Lap}(\frac{\Delta\mathcal{F}_\mathtt{a}}{\epsilon_\mathtt{a}})$$

The noise is added to the output of the learning algorithm $\mathcal{F}(.)$ i.e., the model parameters. Since $\mathcal{K}$ is sampled from the same noise distribution,

$$\mathtt{Lap}(\frac{\Delta\mathcal{F}_\mathtt{c}}{\epsilon_\mathtt{c}}) = \mathtt{Lap}(\frac{\Delta\mathcal{F}_\mathtt{a}}{\epsilon_\mathtt{a}}) \qquad \therefore \frac{\Delta\mathcal{F}_\mathtt{c}}{\epsilon_\mathtt{c}} = \frac{\Delta\mathcal{F}_\mathtt{a}}{\epsilon_\mathtt{a}} \tag{6}$$

From Lemma 1, $\Delta\mathcal{F}_c \leq \Delta\mathcal{F}_a$ and hence $\epsilon_\mathtt{c} \leq \epsilon_\mathtt{a}$. □

While we prove the general result above, our central claim comparing differential privacy for causal and associational models also holds for mechanisms that provide a tighter data-dependent differential privacy guarantee (Papernot et al., 2017). The key idea is to produce an output label based on voting from $\mathtt{M}$ teacher models, each trained on a disjoint subset of the training data. We state the theorem below and provide its proof in Suppl. Section C. Given datasets from different domains, the below theorem also provides a *constructive* proof to train a differentially private causal algorithm following the method from Papernot et al. (2017).

Theorem 4. *Let* $\mathtt{D}$ *be a 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* $\mathtt{n_j}$ *be the votes for the jth class from* $\mathtt{M}$ *teacher models. Let* $\mathcal{M}$ *be the mechanism that produces a noisy max,* $\arg\max_\mathtt{j}\{\mathtt{n_j} + \mathtt{Lap}(2/\gamma)\}$. *Then the privacy budget* $\epsilon_\mathtt{c}$ *for a causal model trained on* $\mathtt{D}$ *is lower than that for an associational model with the same accuracy.*

3.2 Robustness to Membership Attacks

Deep learning models have been shown to memorize or overfit on the training data during the learning process (Carlini et al., 2018). Such overfitted models are susceptible to *membership inference attacks* that can accurately predict whether a target input belongs to the training dataset or not (Shokri et al., 2017). There are multiple variants of the attack depending on the information accessible to the adversary. An adversary with black-box access to a model observes confidence scores for the predicted output whereas one with the white-box access observes all model parameters and the output at each layer in the model (Nasr et al., 2018a). In the black-box setting, a membership attack is possible whenever the distribution of output scores for training data is different from the test data, and has been connected to model overfitting (Yeom et al., 2018). Alternatively, if the adversary knows the distribution of the training inputs, they may learn a "shadow" model based on synthetic inputs and use the shadow model's output to build a membership classifier (Shokri et al., 2017). For the white-box setting, if an adversary knows the true label for the target input, then they may guess membership of the input based on either the loss or gradient values during inference (Nasr et al., 2018a).

Most of the existing membership inference attacks have been demonstrated for test inputs from the same data distribution as the training set. When test inputs are expected from the same distribution, methods to reduce overfitting (such as adversarial regularization) can help reduce privacy risks (Nasr et al., 2018b). However in practice, this is seldom the case. For instance, in our example of a model trained with a single hospital's data, the test inputs may come from different hospitals. Therefore, models trained to reduce the generalization error for a specific test distribution are still susceptible to membership inference when the distribution of features is changed. This is due to the problem of *covariate shift* that introduces a domain adaptation error term (Mansour et al., 2009). That is, the loss-minimizing model that predicts $\mathtt{Y}$ changes with a different distribution, and thus allows the adversary to detect differences in losses for the test versus training datasets. As we show below, causal models alleviate the risk of membership inference attacks. Based on Yeom et al. (2018), we first define a membership attack.

Definition 6. *Let* $\mathtt{h}$ *be trained on a dataset* $\mathtt{S(X,Y)} \sim \mathtt{P}$ *of size* $\mathtt{N}$. *Let* $\mathcal{A}$ *be an adversary with access to* $\mathtt{h}$ *and an input* $\mathtt{x} \sim \mathtt{P}^*$ *where* $\mathtt{P}^*$ *is any distribution such that* $\mathtt{P(Y|X_C)} = \mathtt{P^*(Y|X_C)}$. *Then advantage of an adversary in membership inference is the difference between true and false positive rate in guessing whether the the input belongs to the training set.* $\mathtt{Adv}(\mathcal{A},\mathtt{h},\mathtt{N},\mathtt{P},\mathtt{P}^*) = \Pr[\mathcal{A} = \mathtt{1}|\mathtt{b} = \mathtt{1}] - \Pr[\mathcal{A} = \mathtt{1}|\mathtt{b} = \mathtt{0}]$, *where* $b = 1$ *if the input is in the training set and else is* 0.

As a warmup, we demonstrate the relationship between membership advantage and out-of-distribution generalization using a specific adversary that predicts membership for an input based on the model's loss. This adversary is motivated by empirical membership inference algorithms (Shokri et al., 2017; Nasr et al., 2018a).

Definition 7. *[From (Yeom et al., 2018)] Assume that the loss* L *is bounded by* $\mathtt{B} \in \mathbb{R}^+$. *Then for a model* $\mathtt{h}$ *and an input* $\mathtt{x}$, *a Bounded-Loss adversary* $\mathcal{A}_\mathtt{BL}$ *predicts membership in a train set with probability* $\mathtt{1} - \mathtt{L_x(h,y)}/\mathtt{B}$.

Theorem 5. *Assume a training set* $\mathtt{S}$ *of size* $\mathtt{n}$ *and a loss function* $\mathtt{L}$ *that is bounded by* $\mathtt{B} \in \mathbb{R}^+$. *Under the conditions of Theorem 1 and for a Bounded-Loss adversary* $\mathcal{A}_\mathtt{BL}$, *the worst-case membership advantage of a causal model* $\mathtt{h}^\mathtt{min}_\mathtt{c,S}$ *is lower than that of an associational model* $\mathtt{h}^\mathtt{min}_\mathtt{a,S}$.

$$\max_{\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}_\mathtt{BL}, \mathtt{h}^\mathtt{min}_\mathtt{c,S}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*) \leq \max_{\mathtt{P}^*} \mathtt{Adv}(\mathcal{A}_\mathtt{BL}, \mathtt{h}^\mathtt{min}_\mathtt{a,S}, \mathtt{n}, \mathtt{P}, \mathtt{P}^*)$$

Proof. Let the variable $\mathtt{b} = \mathtt{1}$ denote that data point belongs to the train dataset $\mathtt{S}$. The membership advantage of the bounded loss adversary $\mathcal{A}$ for any model $\mathtt{h}$ trained on dataset $\mathtt{S} \sim \mathtt{P}$ is given by,

$$\begin{aligned}\mathtt{Adv}(\mathcal{A},\mathtt{h},\mathtt{n},\mathtt{P},\mathtt{P}^*) &= \Pr[\mathcal{A}=1|\mathtt{b}=1]-\Pr[\mathcal{A}=1|\mathtt{b}=0]\\ &= \Pr[\mathcal{A}=0|\mathtt{b}=0]-\Pr[\mathcal{A}=0|\mathtt{b}=1]\\ &= \mathbb{E}[\frac{\mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{y})}{\mathtt{B}}|\mathtt{b}=0]-\mathbb{E}[\frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{y})}{\mathtt{B}}|\mathtt{b}=1]\\ &= \frac{1}{\mathtt{B}}(\mathbb{E}_{\mathtt{x}'\sim\mathtt{P}^*}[\mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{y})]-\mathbb{E}_{\mathtt{x}\sim\mathtt{S}}[\mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{y})])\\ &\leq \max_{\mathtt{x}'\notin\mathtt{S}}\mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h},\mathtt{y})\end{aligned}$$

where the third equality is due to Def. 7 for $\mathcal{A}_{\mathtt{BL}}$, and the last inequality is due to the fact that the expected value of a random variable is less than or equal to the maximum value. Note that the upper bound in the above inequality is tight: it can be achieved by evaluating membership advantage only on those $\mathtt{x}'$ that lead to the maximum loss difference. Thus,

$$\max_{\mathtt{P}^*}\mathtt{Adv}(\mathcal{A},\mathtt{h},\mathtt{N},\mathtt{P},\mathtt{P}^*) = \max_{\mathtt{x}'}\mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h},\mathtt{y}) \quad (7)$$

Applying Eqn. 7 to the trained causal model $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}$ and associational model $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}$, we obtain:

$$\max_{\mathtt{P}^*}\mathtt{Adv}(\mathcal{A},\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{n},\mathtt{P},\mathtt{P}^*) = \max_{\mathtt{x}'}\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})$$
$$\max_{\mathtt{P}^*}\mathtt{Adv}(\mathcal{A},\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{n},\mathtt{P},\mathtt{P}^*) = \max_{\mathtt{x}'}\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})$$

From Theorem 2 proof (Suppl. Eqn. 58), we state the inequality, $\max_{\mathtt{x}'}\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y}) \leq \max_{\mathtt{x}'}\mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})$. Combining this inequality with the above equations, we get the main result.

$$\max_{\mathtt{P}^*}\mathtt{Adv}(\mathcal{A},\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{n},\mathtt{P},\mathtt{P}^*) \leq \max_{\mathtt{P}^*}\mathtt{Adv}(\mathcal{A},\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{n},\mathtt{P},\mathtt{P}^*)$$

□

We now prove a more general result. The maximum membership advantage for a causal DP mechanism (based on a causal model) is not greater than that of an associational DP mechanism. We present a lemma from Yeom et al. (2018).

Lemma 2. *[From (Yeom et al., 2018)] Let $\mathcal{M}$ be a ϵ-differentially private mechanism based on a model h. The membership advantage is bounded by* $\exp(\epsilon)-1$.

Based on the above lemma and Theorem 3, we can show that the upper bound of membership advantage for an ϵ_c-DP mechanism from a causal model $e^{\epsilon_c}-1$ is not greater than that of an ϵ_a-DP mechanism from an associational model, $e^{\epsilon_a}-1$, since $\epsilon_c \leq \epsilon_a$. The next theorem proves that the same holds true for the *maximum* membership advantage.

Theorem 6. *Under the conditions of Theorem 1, let* $\mathtt{S} \sim \mathtt{P}(\mathtt{X},\mathtt{Y})$ *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* $\mathtt{P}(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_{\mathtt{C}})$. *Then, across all adversaries* $\mathcal{A}$ *that predict membership in* $\mathtt{S} \sim \mathtt{P}$, *the worst-case membership advantage of* $\hat{\mathcal{F}}_{\mathtt{c,S}}$ *is not greater than that of* $\hat{\mathcal{F}}_{\mathtt{a,S}}$.

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

Proof. We construct an expression for the maximum membership advantage for any ϵ-DP model and then show that it is an increasing function of the sensitivity, and thus ϵ. □

Finally, we show that membership advantage against a causal model trained on infinite data will be zero for any adversary. The proof is based on the result from Theorem 1 that $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P^*}}$ for a causal model. Crucially, membership advantage does not go to zero as $\mathtt{n} \to \infty$ for associational models, since $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}$ in general. Detailed proof is in Suppl. Section E.

Corollary 2. *Under the conditions of Theorem 1, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}$ *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* $\mathtt{Adv}(\mathcal{A},\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) \to 0$.

3.3 Robustness to Attribute Inference Attacks

We prove similar results on the benefits of causal models for attribute inference attacks where a model may reveal the value of sensitive features of a test input, given partial knowledge of its features. For instance, given a model's output and certain features about a person, an adversary may infer other attributes of the person (e.g., their demographics or genetic information). As another example, it can be possible to infer a person's face based on hill climb on the output score of a face detection model (Fredrikson et al., 2015). Model inversion is not always due to a fault in learning: a model may learn a true, generalizable relationship between features and the outcome, but still be vulnerable to a model inversion attack. This is because given $k-1$ features and the true outcome label, it is possible to guess the kth feature by brute-force search on output scores generated by the model.

However, inversion based on learning correlations between features and the outcome, e.g., using demographics to predict disease, can be alleviated by causal models, since a non-causal feature will not be included in the model.

Definition 8 (From (Yeom et al., 2018))**.** *Let* $\mathtt{h}$ *be a model trained on a dataset* $\mathtt{S}$*(X,Y). Let* $\mathcal{A}$ *be an adversary with access to h, and a partial test input* $\mathtt{x}_{\mathtt{A}} \subset \mathtt{x}$. *The attribute advantage of the adversary is the difference between true and false positive rates in guessing the value of a sensitive*

Dataset	**Child**	**Alarm**	**(Sachs)**	**Water**
Output	XrayReport	BP	Akt	CKNI_12_45
No. of classes	5	3	3	3
Nodes	20	37	11	32
Arcs	25	46	17	66
Parameters	230	509	178	10083

Table 1: Details of the benchmark datasets.

feature $\mathtt{x_s} \notin \mathtt{x_A}$. *For a binary* $\mathtt{x_s}$,

$$\mathtt{Adv}(\mathcal{A}, \mathtt{h}) = \mathtt{Pr}(\mathcal{A} = \mathtt{1} | \mathtt{x_s} = \mathtt{1}) - \mathtt{Pr}(\mathcal{A} = \mathtt{1} | \mathtt{x_s} = \mathtt{0})$$

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

The proof is in Suppl. Section F.

4 Implementation and Evaluation

We perform our evaluation on two types of datasets: 1) Four datasets generated from known Bayesian Networks and 2) Colored images of digits from the MNIST dataset.

Bayesian Networks. To avoid errors in learning causal structure from data, we perform evaluation on datasets for which the causal structure and the true conditional probabilities of the variables are known from prior research. We select 4 Bayesian network datasets— Child, Sachs, Alarm and Water that range from 178-10k parameters (Table 1)[2]. Nodes represent the number of input features and arcs denote the causal connections between these features in the network. Each causal connection is specified using a conditional probability table $\mathtt{P}(\mathtt{X_i} | \mathtt{Parents}(\mathtt{X_i}))$; we consider these probability values as the parameters in our models. To create a prediction task, we select a variable in each of these networks as the output Y. The number of classes in Table 1 denote the possible values for an output variable. For example, the variable BP (blood pressure) in the alarm dataset takes 3 values i.e, `LOW, NORMAL, HIGH`. The causal model uses only parents of Y whereas the associational model (DNN) uses all nodes except Y as features.

Colored MNIST Dataset. We also evaluate on a complex dataset where it is difficult to construct a causal graph of the input features. For this, we consider the dataset of colored MNIST images used in a recent work by (Arjovsky et al., 2019). The original MNIST dataset consists of grayscale images of handwritten digits (0-9)[3]. The colored MNIST dataset consists of inputs where digits 0-4 are red in color with label as 0 while 5-9 are green in color and have label 1. The training dataset consists of two environments where only 10% and 20% of inputs *do not* follow the correlation of color to digits. This creates a spurious correlation of color with the output. In this dataset, *shape* of the digit is the actual causal feature whereas *color* acts as the associational or non-causal feature. The test dataset is generated such that 90% of the inputs *do not* follow the color pattern. We use the code made available by (Arjovsky et al., 2019) to generate the dataset and perform our evaluation[4]. We refer the readers to the paper for further details.

4.1 Results for Bayesian Networks Dataset

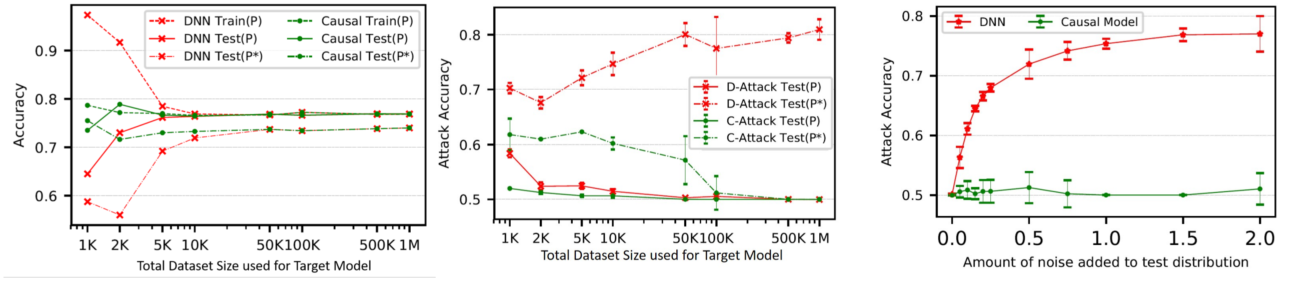
Evaluation Methodology. We sample data using the causal structure and probabilities from the Bayesian network, and use a 60:40% split for train-test datasets. We learn a causal model and a deep neural network (DNN) on each training dataset. We implement the attacker model to perform membership inference attack using the output confidences of both these models, based on past work (Salem et al., 2018). The input features for the attacker model comprises of the output confidences from the target model, and the output is membership prediction (member / non-member) in the training dataset of the target model. In both the train and the test data for the attacker model, the number of members and non-members are equal. The creation of the attacker dataset is described in Figure 5 in the Suppl. Section G. Note that the attack accuracies reported are an upper bound since we assume that the adversary has access to the subset of training data for the ML model.

To train the causal model, we use the bnlearn library in R language that supports maximum likelihood estimation of the parameters in Y's conditional probability table. For prediction, we use the `parents` method to predict the class of any specific variable. To train the DNN model and the attacker model, we build custom estimators in Python using Tensorflow v1.2. The DNN model is a multilayer perceptron (MLP) with 3 hidden layers of 128, 512 and 128 nodes respectively. The learning rate is set to 0.0001 and the model is trained for 10000 steps. The attacker model has 2 hidden layers with 5 nodes each, a learning rate of 0.001, and is trained for 5000 steps. Both models use Adam optimizer, ReLU for the activation function, and cross entropy as the loss function. We chose these parameters to ensure model convergence. We evaluate the DNN and the causal model sample sizes ranging from 1K to 1M dataset sizes. We refer Test(P) as the test dataset which is drawn from the same distribution as the training data and Test(P*) is generated from a completely different distribution except for the relationship of the output class to its parents. To generate Test(P*), we alter the true probabilities $\mathtt{Pr}(\mathtt{X})$ uniformly at random (later, we consider adding noise to the original value). Our goal with generating Test (P*) is to capture extreme shifts in distribution for input features. The causal and DNN model are the *target* on which the attack is perpetrated.

[2] www.bnlearn.com/bnrepository

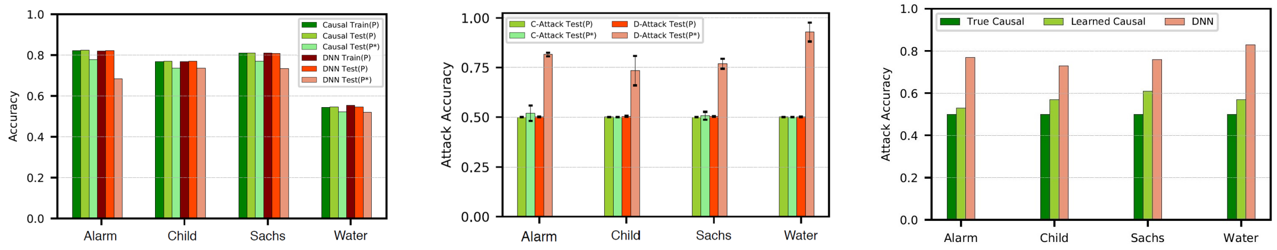
[3] http://yann.lecun.com/exdb/mnist/

[4] https://github.com/facebookresearch/InvariantRiskMinimization



(a) (b) (c)

Figure 3: Results for Child dataset with XrayReport as the output. (a) is the target model accuracy. (b) is the attack accuracy for different dataset sizes on which the target model is trained and (c) is the attack accuracy for test distribution with varying amount of noise for total dataset size of 100K samples.



(a) (b) (c)

Figure 4: Results for all the bayesian models trained on dataset of size of 60K. (a) is the accuracy of the target model, (b) is the attack accuracy for the target model, (c) is the attack accuracy using Test(P*) dataset on true causal, learned causal and DNN models.

Accuracy comparison of DNN and Causal models. Figure 3a shows the target model accuracy comparison for the DNN and the causal model trained on the Child dataset with XrayReport as the output variable. We report the accuracy of the target models only for a single run since in practice the attacker would have access to the outputs of only a single model. We observe that the DNN model has a large difference between the train and the test accuracy (both Test(P) and Test(P*)) for smaller dataset sizes (1K and 2K). This indicates that the model overfits on the training data for these dataset sizes. However, after 10K samples, the model converges such that the train and Test(P) dataset have the same accuracy. The accuracy for the Test(P*) distribution stabilizes for a total dataset size of 10K samples. In contrast, for the causal model, the train and Test(P) accuracy are similar for the causal model even on smaller dataset sizes. However, after convergence at around 10K samples, the gap between the accuracy of train and Test(P*) dataset is the same for both the DNN and the causal model. Figure 4a shows similar results for the accuracy on all the datasets.

Attack Accuracy of DNN and Causal models. A naive attacker classifier would predict all the samples to be members and therefore achieve 0.5 prediction accuracy. Thus, we consider 0.5 as the baseline attack accuracy which is equal to a random guess. Figure 3b shows the attack accuracy comparison for Test(P) (same distribution) and Test(P*) (different distribution) datasets. Attack accuracy of the Test(P) dataset for the causal model is slightly above a random guess for smaller dataset sizes, and then converges to 0.5. In comparison, attack accuracy for the DNN on Test(P) dataset is over 0.6 for smaller samples sizes and reaches 0.5 after 10K datapoints. This confirms past work that an overfitted DNN is susceptible to membership inference attacks even for test data generated from the same distribution as the training data (Yeom et al., 2018). On Test(P*), the attack accuracy is always higher for the DNN than the causal model, indicating our main result that associational models “overfit” to the training distribution, in addition to the training dataset. Membership inference accuracy for DNNs is as high as 0.8 for total dataset size of 50K while that of causal models is below 0.6. Further, attack accuracy for DNN increases with sample size whereas attack accuracy for the causal model reduces to 0.5 for total dataset size over 100k even when the gap between the train and test accuracies is the same as DNNs (Figure 3a). These results show that causal models generalize better than DNNs across input distributions. Figure 4b shows a similar result for all four datasets. The attack accuracy for DNNs and the causal model is close to 0.5 for the Test 1 dataset while for the Test(P*) dataset the attack accuracy is significantly higher for DNNs than causal model. This empirically confirms our claim that in general, causal models are robust to membership inference attacks across test distributions as compared to associational models.

Attack Accuracy for Different Test Distributions. To understand the change in attack accuracy as $\Pr(X)$ changes, we generate test data from different distributions by adding varying amount of noise to the true probabilities. We range the noise value between 0 to 2 and add it to the individual

	True Model	Learned Causal (2 causal +)		DNN
Acc. (%)	2 causal parents	1 non-causal parent	2 non-causal parents	
Attack	**50**	**52**	**61**	**76**
Pred.	79	75	68.8	73

Table 2: Attack and Prediction accuracy comparison across models for `Sachs` dataset and `Akt` output variable.

probabilities which are then normalized to sum up to 1. Figure (3c) shows the attack accuracy for the causal model and the DNN on the child dataset for a total sample size of 100K samples. We observe that the attack accuracy increases with increase in the noise values for the DNN. Even for a small amount of noise, attack accuracies increase sharply. In contrast, attack accuracies stay close to 0.5 for the causal model, demonstrating the robustness to membership attacks.

Results with learnt causal model. Finally, we perform experiments to understand the effect of privacy guarantees on causal structures learned from data that might vary from the true causal structure. For these datasets, a simple hill-climbing algorithm returned the true causal parents. Hence we evaluated attack accuracy for models with hand-crafted errors in learning the structure, i.e., misestimation of causal parents, see Figure (4c). Specifically, we include two non-causal features as parents of the output variable along with the true causal features. The attack risk increases as a learnt model deviates from the true causal structure, however it still exhibits lower attack accuracy than the corresponding associational model. Table 2 shows the attack and prediction accuracy for `Sachs` dataset when trained with increase in error in the causal model (with 1 and 2 non-causal features), and the results for the corresponding DNN model.

4.2 Results for Colored MNIST Dataset

In recent work Arjovsky et al. (2019) proposed a way to train a causal model by minimizing the risk across different environments or distributions of the dataset. Using this approach, we train an invariant risk minimizer (IRM) and an emprical risk minimizer (ERM) on the colored MNIST data. Since IRM constructs the same model using invariant feature representation for the two training domains, it is aimed to learn the causal features (shape) that are also invariant across domains (Peters et al., 2016). Thus IRM can be considered as a causal model while ERM is an associational model. Table 3 gives the model accuracy and the attack accuracy for IRM and ERM models. The attacker model has 2 hidden layers with 3 nodes each, a learning rate of 0.001, and is trained for 5000 steps. We observe that the causal model has attack accuracy close to a random guess while the associational model has 66% attack accuracy. Although the training accuracy of IRM is lower than ERM, we expect this to be an acceptable trade-off for the stronger privacy and better generalizability guarantees of causal models.

Model	Train Acc. (%)	Test Acc. (%)	Attack Acc. (%)
IRM (causal)	70	69	53
ERM (Associational)	87	16	66

Table 3: Results on Colored MNIST Dataset.

5 Related Work

Privacy attacks and defenses on ML models. Shokri et al. (2017) demonstrate the first membership inference attacks on black box neural network models with access only to the confidence values. Similar attacks have been shown on several other models such as GANs (Hayes et al., 2017), text prediction generative models (Carlini et al., 2018; Song & Shmatikov, 2018) and federated learning models (Nasr et al., 2018b). However, prior research does not focus on the severity of these attacks with change in the distribution of the test dataset. We discussed in Section 3.2 that existing defenses based on regularization (Nasr et al., 2018b) are not practical when models are evaluated on test inputs from different distributions. Another line of defense is to add differentially private noise while training the model. However, the ϵ values necessary to mitigate membership inference attacks in deep neural networks require addition of large amount of noise that degrades the accuracy of the output model (Rahman et al., 2018). Thus, there is a trade-off between privacy and utility when using differential privacy for neural networks. In contrast, we show that causal models require lower amount of noise to achieve the same ϵ differential privacy guarantees and hence retain accuracy closer to the original model. Further, as training sample sizes become sufficiently large (Section 4) causal models are robust to membership inference attacks across distributions.

Causal learning and privacy. There is substantial literature on learning causal models from data; for a review see (Peters et al., 2017; Pearl, 2009). Kusner et al. (2015) proposed a method to privately reveal parameters from a causal learning algorithm, using the framework of differential privacy. Instead of a specific causal algorithm, our focus is on the privacy benefits of causal models for general predictive tasks. While recent work uses causal models to study properties of ML models such as providing explanations (Datta et al., 2016) or fairness (Kusner et al., 2017), the relation of causal learning to model privacy is yet unexplored.

6 Conclusion and Future Work

Our results show that causal learning is a promising approach to train models that are robust to privacy attacks such as membership inference and model inversion. As future work, we aim to investigate privacy guarantees when the causal features and the relationship between them is not known apriori and with causal insufficiency and selection bias in the observed data.

Acknowledgements

We thank the reviewers for their useful comments on the paper. We thank Olga Ohrimenko, Boris Koepf, Amit Deshpande and Emre Kiciman for helpful discussion and feedback on this work.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan):171–234, 2010.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., and Song, D. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232*, 2018.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pp. 598–617. IEEE, 2016.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

Fischer, T. and Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333. ACM, 2015.

Hamm, J., Cao, Y., and Belkin, M. Learning privately from multiparty data. In *International Conference on Machine Learning*, pp. 555–563, 2016.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Kusner, M. J., Sun, Y., Sridharan, K., and Weinberger, K. Q. Private causal inference. *arXiv preprint arXiv:1512.05469*, 2015.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, pp. 1931. NIH Public Access, 2018.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. *arXiv preprint arXiv:1812.00910*, 2018a.

Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646. ACM, 2018b.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pellet, J.-P. and Elisseeff, A. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(Jul):1295–1342, 2008.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Rahman, M. A., Rahman, T., Laganiere, R., Mohammed, N., and Wang, Y. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 2018.

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Scutari, M. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 3–18. IEEE, 2017.

Song, C. and Shmatikov, V. The natural auditor: How to tell if someone used your words to train their model. *arXiv preprint arXiv:1811.00513*, 2018.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. Using deep learning to detect price change indications in financial markets. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2511–2515. IEEE, 2017.

Wu, X., Fredrikson, M., Wu, W., Jha, S., and Naughton, J. F. Revisiting differentially private regression: Lessons from learning theory and their consequences. *arXiv preprint arXiv:1512.06388*, 2015.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Supplementary Material: Alleviating Privacy Attacks via Causal Learning

A Generalization Properties of Causal Models

A.1 Generalization over Different Distributions

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

Theorem 1. *Consider a structural causal graph* $\mathtt{G}$ *that connects* $\mathtt{X}$ *to* $\mathtt{Y}$*, and causal features* $\mathtt{X_C} \subset \mathtt{X}$ *where* $\mathtt{X_C}$ *represent the parents of* $\mathtt{Y}$ *under* $\mathtt{G}$*. Let* $\mathtt{P}(\mathtt{X},\mathtt{Y})$ *and* $\mathtt{P}^*(\mathtt{X},\mathtt{Y})$ *be two distributions with arbitrary* $\mathtt{P}(\mathtt{X})$ *and* $\mathtt{P}^*(\mathtt{X})$*, having overlap,* $\mathtt{P}(\mathtt{X}=\mathtt{x}) > \mathtt{0}$ *whenever* $\mathtt{P}^*(\mathtt{X}=\mathtt{x}) > \mathtt{0}$*. In addition, the causal relationship between* $\mathtt{X_C}$ *and* $\mathtt{Y}$ *is preserved, which implies that* $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$*. Let* L *be a symmetric loss function that obeys the triangle inequality (such as L1, L2 or 0-1 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* $\Pr(\mathtt{Y}|\mathtt{X_C})$ *(e.g.,* $\mathtt{f} := \mathbb{E}[\mathtt{Y}|\mathtt{X_C}]$ *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* $\mathtt{Y}|\mathtt{X_C}$ *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* $\mathtt{S}\sim\mathtt{P}(\mathtt{X},\mathtt{Y})$ *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* $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}$ *on* $\mathtt{P}$ *and* $\mathtt{P}^*$ *respectively.*

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

   However, 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 hypotheses (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 $\mathtt{S}\sim\mathtt{P}(\mathtt{X},\mathtt{Y})$. 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,

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

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, $\mathcal{H}_{\mathcal{C}}$ refers to the class of models that uses all causal features ($\mathtt{X}_{\mathtt{C}}$), 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. $\mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{y})$ 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:

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

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 P AND P*

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

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

$$\Pr(\mathtt{Y}|\mathtt{do}(\mathtt{X}_{\mathtt{c}}=\mathtt{x}_{\mathtt{c}})) = \mathtt{P}(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}=\mathtt{x}_{\mathtt{c}}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}=\mathtt{x}_{\mathtt{c}}) \tag{16}$$

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

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

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

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

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

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

Associational Model. In contrast, an associational model may use a subset $\mathtt{X}_{\mathtt{A}} \subseteq \mathtt{X}$ 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 $\mathtt{h}_{\mathtt{a},\mathtt{P}}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}_{\mathtt{a}}\in\mathcal{H}_{\mathtt{A}}} \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{a}},\mathtt{y})$ and $\mathtt{h}_{\mathtt{a},\mathtt{P}^*}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}_{\mathtt{a}}\in\mathcal{H}_{\mathcal{A}}} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{a}},\mathtt{y})$, we can write,

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

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 $\mathtt{h^{OPT}_{a,P^*}}$, we can write:

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

Now, in general,

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

Even if the optimal associational model $\mathtt{f_A} \in \mathcal{H}_\mathtt{A}$ (as we assumed for causal models), and thus $\mathtt{f_{P(X_A,Y)}} = \mathtt{f_{P(Y|X_A)}}$ and $\mathtt{f_{P^*(X_A,Y)}} = \mathtt{f_{P^*(Y|X_A)}}$, they are not the same since $\mathtt{P(Y|X_A)} \neq \mathtt{P^*(Y|X_A)}$. Therefore we obtain,

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

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

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, $\mathcal{L}_\mathtt{P}(\mathtt{h^{OPT}_{c,P}},\mathtt{y}) = \mathcal{L}_{\mathtt{P^*}}(\mathtt{h^{OPT}_{c,P^*}},\mathtt{y}) = \mathtt{0}$. Similarly, $\mathcal{L}_\mathtt{P}(\mathtt{h^{OPT}_{a,P}},\mathtt{y}) = \mathtt{0}$. (Note that here we consider only those cases where $\mathtt{f_{P(Y|X)}} \in \mathcal{H}_\mathtt{A}$ and $\mathtt{f_{P^*(Y|X)}} \in \mathcal{H}_\mathtt{A}$ for a fair comparison; otherwise, the error bound for $\mathtt{h_a} \in \mathcal{H}_\mathtt{A}$ is trivially larger than that for $\mathtt{h_c} \in \mathcal{H}_\mathtt{C}$).

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

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

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

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

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

For an associational model, we obtain,

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

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 $1-\delta$:

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

Similarly, for the associational model,

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

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

$$\mathtt{ODE\text{-}Bound_{P,P^*}}(\mathtt{h_c},\mathtt{y};\delta) \leq \mathtt{ODE\text{-}Bound_{P,P^*}}(\mathtt{h_a},\mathtt{y};\delta) \tag{28}$$

IIIb. CLAIM 2

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

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

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

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

Similarly, for associational model,

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

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, $\mathcal{L}_\mathtt{P}(\mathtt{h^{OPT}_{a,P}},\mathtt{y}) \leq \mathcal{L}_\mathtt{P}(\mathtt{h^{OPT}_{c,P}},\mathtt{y})$.

However, since $\mathtt{h^{OPT}_{a,P}} \neq \mathtt{h^{OPT}_{a,P^*}}$, 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.

$$\begin{aligned} \gamma_1 =& \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y}) && (32)\\ \gamma_2 =& \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y}) - \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y}) && (33)\end{aligned}$$

Then, Eqn. 31 transforms to,

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

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

$$\mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c}, \mathtt{y}; \delta) \leq \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a}, \mathtt{y}; \delta) \tag{35}$$

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

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

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

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

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

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

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

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

From condition (ii) above, $\mathbb{E}_{\mathtt{P}^*}[\lambda_{\mathtt{x_A}}] \geq \mathbb{E}_{\mathtt{P}}[\lambda_{\mathtt{x_A}}]$, 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 $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}$ as adding some arbitrary value $\lambda^{(\mathtt{i})}_{\mathtt{x_A}}$ to $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}(\mathtt{x}_{\mathtt{c}}^{(\mathtt{i})})$; and then construct a $\mathtt{P}^*$ such that the relationship $\Pr(\mathtt{Y}|\mathtt{X_A})$ is reversed on $\mathtt{P}^*$, and thus $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}^*}$ 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 $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y}; \lambda) - \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}^*}, \mathtt{y}) \geq \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y}; \lambda)$.

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

$$\mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c}, \mathtt{y}; \delta) \leq \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a}, \mathtt{y}; \delta) \tag{41}$$

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 $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y})$ and $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y})$. Since γ_2 increases monotonically with $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{y})$ ($\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y})$ is bounded by $\max_{\mathtt{x}} \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{y})$), 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,

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

□

A.2 Generalization over a Single Datapoint

Theorem 2. *Consider a causal model* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} : \mathtt{X_C} \to \mathtt{Y}$ *and an associational model* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}} : \mathtt{X} \to \mathtt{Y}$ *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,* $\mathtt{y} \sim \mathtt{P}(\mathtt{Y}|\mathtt{X_C} = \mathtt{x})$ *and* $\mathtt{y}' \sim \mathtt{P}(\mathtt{Y}|\mathtt{X_C} = \mathtt{x}')$*. Then, the worst-case generalization error for a causal model on such* $\mathtt{x}'$ *is less than or equal to that for an associational model.*

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

Proof. For any model 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 $x \in S$, $\min_{\mathtt{x}\in\mathtt{S}} \mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{y})$ is lower for the associational model than a causal model.

Loss on a training data point. First, consider loss on $\mathtt{x} \in \mathtt{S}$, $\mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{y})$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where $\mathtt{h_a}$ is some associational function of $\mathtt{x}$. Therefore the loss on $\mathtt{x}'_1$ becomes,

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

Since $\Pr(\mathtt{Y}|\mathtt{X_A})$ can change for different $\mathtt{x}' \sim \mathtt{P}^*$ (where $\mathtt{X_A} = \mathtt{X} \setminus \mathtt{X_C}$ 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,

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

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

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

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

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

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

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

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

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

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

Proof. Let $S_{n-1} = S \setminus (x, y))$ and similarly $S'_{n-1} = S' \setminus (x', y')$. Since S and 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 S_{n-1}, $(n-1)L_{S_{n-1}}$ to Theorem 2 statement.

Considering the LHS of Theorem 2,

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

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

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

Using Theorem 2 and dividing Eqns. 59 and 60 by n, we obtain,

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

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

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

□

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

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

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

Using the definition of $\mathcal{L}_S = \frac{1}{n}\sum_{i=1}^{n} L_{x_i}(h, y)$, we can write the following for any two neighboring datasets 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 $\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S'}},\mathtt{y})-\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{S}},\mathtt{y})\le\mathtt{0}$.

□

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* $\theta\in\Omega\subseteq\mathbb{R}^n$. *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* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S'}}\leftarrow\mathcal{F}_{\mathtt{a}}(\mathtt{S'})$,

$$\Delta\mathcal{F}_{\mathtt{c}}=\max_{\mathtt{S,S'}}||\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}}||_{\mathtt{1}}\le\max_{\mathtt{S,S'}}||\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S'}}||_{\mathtt{1}}=\Delta\mathcal{F}_{\mathtt{a}}$$

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 $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}}$ trained on $\mathtt{S}$ and $\mathtt{S'}$ respectively (Wu et al., 2015),

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

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

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

Now consider $\max_{\mathtt{S,S'}}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}}\right\|_{\mathtt{1}}$. Without loss of generality, we can order the pair of datasets $\mathtt{S},\mathtt{S'}$ such that $\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}},\mathtt{y})\le\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})$. Then using Eqn. 66 and taking the maximum, we obtain,

$$\begin{aligned}
\frac{\lambda}{\mathtt{2}}\max_{\mathtt{S,S'}}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}}\right\|_{\mathtt{1}}^{\mathtt{2}}&\le\max_{\mathtt{S,S'}}\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S'}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&\le\max_{\mathtt{S,S'}}\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}},\mathtt{y})\\
&\le\max_{\mathtt{S,S'}}\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})
\end{aligned}\tag{67}$$

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 $\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})\ge\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})$ since $\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}$ is the loss-minimizing model over S'_1. Therefore, we can rewrite,

$$\begin{aligned}
&\max_{\mathtt{S,S'}}\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
&=\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})\\
&\le\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})+(\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})-\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}))\\
&=[\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})]+(\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})]
\end{aligned}\tag{68}$$

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

$$\begin{aligned}
&\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})\\
&\le\frac{\mathtt{L}_{\mathtt{x'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathtt{L}_{\mathtt{x'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})}{\mathtt{n}}+\frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})-\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})}{\mathtt{n}}\\
&\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})\\
&\le\frac{\mathtt{L}_{\mathtt{x'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathtt{L}_{\mathtt{x'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})}{\mathtt{n}}+\frac{\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})-\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})}{\mathtt{n}}
\end{aligned}\tag{69}$$

Now since the loss function $\mathtt{L}(.,\mathtt{y})$ is ρ-Lipschitz, we have $\mathtt{L}_{\mathtt{x}}(\mathtt{h}_1,\mathtt{y})-\mathtt{L}_{\mathtt{x}}(\mathtt{h}_2,\mathtt{y})\le\rho\|\mathtt{h}_1-\mathtt{h}_2\|_{\mathtt{1}}$ for any data point $\mathtt{x}$ and any two models $\mathtt{h}_1$ and $\mathtt{h}_2$. Plugging Eqn. 69 and the ρ-Lipschitz property back in Eqn. 68,

$$\begin{aligned}
&\quad\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})\le\frac{\mathtt{2}}{\mathtt{n}}\rho\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}\right\|_{\mathtt{1}}\\
&\quad\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})\le\frac{\mathtt{2}}{\mathtt{n}}\rho\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}\right\|_{\mathtt{1}}\\
&\Rightarrow\max_{\mathtt{S,S'}}\mathcal{L}_{\mathtt{S'}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})-\mathcal{L}_{\mathtt{S}}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}},\mathtt{y})\\
&\quad\le[\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})-\mathcal{L}_{\mathtt{S}'_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})]+(\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1})-\mathcal{L}_{\mathtt{S}_1}(\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1},\mathtt{y})]\\
&\quad\le\frac{\mathtt{2}}{\mathtt{n}}\rho\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}\right\|_{\mathtt{1}}+\frac{\mathtt{2}}{\mathtt{n}}\rho\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}\right\|_{\mathtt{1}}\\
&\quad=\frac{\mathtt{4}}{\mathtt{n}}\rho\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a,S}'_1}\right\|_{\mathtt{1}}
\end{aligned}\tag{70}$$

Finally, combining with Eqn. 67, we obtain,

$$\begin{aligned}\max_{\mathtt{S},\mathtt{S}'}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}\right\|_{\mathtt{1}}^2 &\le \frac{8\rho}{\lambda\mathtt{n}}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'_1}\right\|_{\mathtt{1}}\\ &\le \frac{8\rho}{\lambda\mathtt{n}}\max_{\mathtt{S},\mathtt{S}'}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}\right\|_{\mathtt{1}} \qquad (71)\\ &\le \max_{\mathtt{S},\mathtt{S}'}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}\right\|_{\mathtt{1}}\end{aligned}$$

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

$$\max_{\mathtt{S},\mathtt{S}'}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}\right\|_{\mathtt{1}} \le \max_{\mathtt{S},\mathtt{S}'}\left\|\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}\right\|_{\mathtt{1}} \qquad (72)$$

Hence, sensitivity of a causal model is lower than an associational model, i.e., $\Delta\mathcal{F}_c \le \Delta\mathcal{F}_a$. □

C Differential Privacy Guarantees with Tighter Data-dependent Bounds

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

As a consequence of Theorem 2, another generalization property of causal learning is that classification models trained on data from two different distributions $\mathtt{P}(\mathtt{X})$ and $\mathtt{P}^*(\mathtt{X})$ are likely to output the same value for a new input.

Lemma 4. *Under the conditions of Theorem 1 and 0-1 loss, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ *be the loss-minimizing causal classification model trained on a dataset* $\mathtt{S}$ *from distribution* $\mathtt{P}$ *and let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}^*}$ *be the loss-minimizing model trained on a dataset* $\mathtt{S}^*$ *from* $\mathtt{P}^*$. *Similarly, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}^*}$ *be loss-minimizing associational classification models trained on* $\mathtt{S}$ *and* $\mathtt{S}^*$ *respectively. Then for any new data input* $\mathtt{x}$,

$$\begin{aligned}&\min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}^*}(\mathtt{x})\big)\\ &\qquad\ge \min_{\mathtt{S}\sim\mathtt{P},\mathtt{S}^*\sim\mathtt{P}^*}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}^*}(\mathtt{x})\big)\end{aligned}$$

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

Proof. Let $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{P}} = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{S}}(\mathtt{h},\mathtt{y})$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{P}^*} = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{S}^*}(\mathtt{h},\mathtt{y})$ 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 $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{P}}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{P}^*}$. Likewise, let $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{y})$ and similarly let $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*} = \arg\min_{\mathtt{h}\in\mathcal{H}_{\mathtt{A}}}\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{y})$ be the loss-minimizing hypotheses over the two distributions. We can analogously define $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}$ and $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$. 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 models on S and S^* approach their loss-minimizing functions on the distributions P and P^* respectively. Then, for any input $\mathtt{x}$,

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

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

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

$$\begin{aligned}\lim_{|\mathtt{S}|\to\infty}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}(\mathtt{x})=\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}^*}(\mathtt{x})\big)&=\mathtt{1} \qquad (75)\\ \ge \lim_{|\mathtt{S}^*|\to\infty}\Pr\big(\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}(\mathtt{x})&=\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}^*}(\mathtt{x})\big) \qquad (76)\end{aligned}$$

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 $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1^*}$.

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

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

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

where $\mathtt{h}_{\mathtt{a},\mathtt{S}_1}:\mathtt{X}\to\{\mathtt{0},\mathtt{1}\}$ and $\mathtt{h}_{\mathtt{a},\mathtt{S}_1^*}:\mathtt{X}\to\{\mathtt{0},\mathtt{1}\}$ are any two functions. Effectively, when $\mathtt{h}_{\mathtt{a},\mathtt{S}_1}$ 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 $\mathtt{h}_{\mathtt{a},\mathtt{S}_1^*}$ 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 $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1^*}=\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_2^*}$ but the loss-minimizing associational model $\mathtt{h}_{\mathtt{a},\mathtt{S}_2^*}$ has the following property. $\mathtt{h}_{\mathtt{a},\mathtt{S}_2^*}\ne\mathtt{h}_{\mathtt{a},\mathtt{S}_1}(\mathtt{x})$ if $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1^*}=\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1}$, and $\mathtt{h}_{\mathtt{a},\mathtt{S}_2^*}=\mathtt{h}_{\mathtt{a},\mathtt{S}_1}(\mathtt{x})$ if $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1^*}\ne\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1}$. Under $\mathtt{S}_2^*$,

$$\begin{aligned}\left|\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}_2^*}\right|&\ge\left|\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_2^*}\right|\\ =\left|\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}_1^*}\right|&=\max_{\mathtt{S},\mathtt{S}^*}\left|\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}-\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}^*}\right|\end{aligned} \qquad (79)$$

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.

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

□

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 $\epsilon_c \le \epsilon$.

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

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'_j be the vote counts for each class under D'. Because the change in a single input can only affect one model's vote, $|n_j - n'_j| \le 1$.

Let the noise added to each class be $r_j \sim Lap(2/\gamma)$. 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',

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

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

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

for some $\epsilon > 0$. As $P(i|D) = P(r_i \ge r^*)$ 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 $\mathtt{S} \sim \mathtt{P}$, $\mathtt{S}^* \sim \mathtt{P}^*$ such that $\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{C})$ is higher for a causal model, and thus $\max_\mathtt{D} \mathtt{P}(\mathtt{r}_\mathtt{i} \ge \mathtt{r}^*)$ 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 $\mathtt{h}_{\mathtt{c},\mathtt{D}_1}$ and $\mathtt{h}_{\mathtt{c},\mathtt{D}_2}$, and the two associational teacher models be $\mathtt{h}_{\mathtt{a},\mathtt{D}_1}$ and $\mathtt{h}_{\mathtt{a},\mathtt{D}_2}$. From Lemma 4, for any new $\mathtt{x}$, there is more consensus among causal models.

$$\begin{aligned}
&\min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{c,D_1}(x)} = \mathtt{h_{c,D_2}(x)}) \geq \min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{a,D_1}(x)} = \mathtt{h_{a,D_2}(x)}) \\
&\Rightarrow \min_{\mathtt{D}} \min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{c,D_1}(x)} = \mathtt{h_{c,D_2}(x)}) \\
&\geq \min_{\mathtt{D}} \min_{\mathtt{D_1,D_2}} \Pr(\mathtt{h_{a,D_1}(x)} = \mathtt{h_{a,D_2}(x)})
\end{aligned}$$

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^*.

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

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.

Thus, 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* $\hat{\mathcal{F}}_{\mathtt{c,S}}$ *is not greater than that of* $\hat{\mathcal{F}}_{\mathtt{a,S}}$.

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

Proof. Consider a neighboring dataset S' to $\mathtt{S} \sim \mathtt{P}$ such that $\mathtt{S'}$ replaces data point $x \in S$ with a different point x'. Let F_S and F_S' be differentially private mechanisms trained on S and S' respectively. 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:

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

where $\mathcal{A}(x, F_S)$ denotes a membership adversary for an algorithm F_S trained on a dataset S, and $\mathcal{A}(\mathtt{x}, \mathtt{F_{S'}})$ denotes an adversary attacking algorithm $F_{S'}$ trained on S'. Without loss of generality for the case where there are an infinite number of models h, assume that models are sampled from a discrete set of K models: $\{h_1, h_2, ..., h_K\}$. Then using the law of total probability over the models yielded by the algorithms F_S and $F_{S'}$,

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

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

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

$$\begin{aligned}
&\max_{\mathcal{A},\mathtt{P^*}} \mathtt{Adv}(\mathcal{A}, \mathtt{F_S}, \mathtt{n}, \mathtt{P}, \mathtt{P^*}) \\
&\quad = \mathtt{max_{S,S'}} \sum_{\mathtt{h_j} \in \mathtt{H_+}} [\mathtt{P(F_S = h_j)} - \mathtt{P(F'_S = h_j)}] \\
&\quad = \mathtt{max_{S,S'}} \mathtt{P(F_S} \in \mathtt{H_+)} - \mathtt{P(F'_S} \in \mathtt{H_+)} \\
&\quad = \mathtt{max_{S,S'}} \mathtt{P(F_S} \in \mathtt{H_+)} - (1 - \mathtt{P(F'_S} \in \mathtt{H_-))}) \\
&\quad = \mathtt{max_{S,S'}} 2 \Pr(\mathtt{F_S} \in \mathtt{H_+}) - \mathtt{1}
\end{aligned} \tag{87}$$

where the last equality is since the Laplace noise is added to F_S and $F_{S'}$ is from identical distributions and thus $\Pr(F_S \in H_+) = \Pr(F_{S'} \in H_-)$. Equation 87 provides the maximum membership advantage for any ϵ-DP mechanism F_S with Laplace noise.

We next show that Eqn. 87 for a causal mechanism F_c is not greater than that for an associational mechanism F_a. Let $\Pr(F_S)$ be a Laplace distribution with mean at $h_{S,min}$ and $\Pr(F_{S'})$ be a Laplace distribution with mean $h_{S',min}$ with identical scale/noise parameter. We would like to find the boundary model $h^\dagger$ of the set H_+ where $P(F_S = h_j) = P(F'_S = h_j)$, since $\Pr(F_S \in H_+)$ is the probability under the Laplace distribution cut off at a point $h^\dagger$. Due to identical noise for F_S and F'_S and the symmetry of the Laplace distribution, the boundary $h^\dagger$ corresponds to the midpoint of $h_{S,min}$ and $h_{S',min}$: $0.5(\mathtt{h_{S,min}} + \mathtt{h_{S',min}})$. Alternatively, the L1 distance of the boundary $h^\dagger$ from the means of the Laplace distributions can be written as (for a worst S, S'),

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

where ΔF_S is the sensitivity of F_S and the last equality is due to the choice of worst-case S and S'.

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

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

Thus, L1 distance of $h_c^\dagger$ from the mean $h_{c,S}^{min}$ is lower for a causal learning function, and thus its PDF $\Pr(F_S = h_j)$ is higher. Now the set H+ is a one-sided boundary on the values of h and includes the mean of the Laplace distribution. Given symmetry of the Laplace distribution, probability of F_S lying in H_+, $\Pr(\mathtt{F_S} \in \mathtt{H}_+)$ should be lower whenever the PDF at the one-sided boundary is higher. Therefore, $P(F_S in H+)$ is lower for a causal mechanism than the associational learning mechanism.

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

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

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

□

E Infinite sample robustness to membership inference attacks

Corollary 2. *Under the conditions of Theorem 1, let* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}$ *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* $\mathtt{Adv}(\mathcal{A}, \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) \to 0$.

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

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

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

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

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

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

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

where the second last equality follows since any function of $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}$ 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* n *and a structural causal model that connects* $\mathtt{X}$ *to* $\mathtt{Y}$, *a causal model* $\mathtt{h_c}$ *makes it impossible to infer non-causal features.*

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

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

□

G Dataset Distribution

The target model is trained using the synthetic training and test data generated using the bnlearn library. We first divide the total dataset into training and test dataset in a 60:40 ratio. Further, the output of the trained model for each of the training and test dataset is again divided into 50:50 ratio.

The training set for the attacker model consists of confidence values of the target model for the training as well as the test dataset. The relation is explained in Figure 5. Note that the attacker model is trained on the confidence output of the target models.

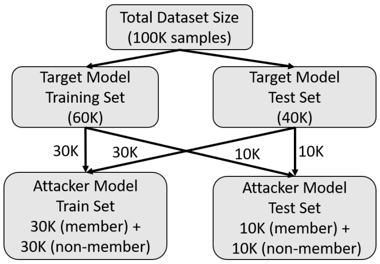


Figure 5: Dataset division for training target and attacker models.