ALLEVIATING PRIVACY ATTACKS VIA CAUSAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models, especially deep neural networks have been shown to reveal membership information of inputs in the training data. Such *membership inference* attacks are a serious privacy concern, for example, patients providing medical records to build a model that detects HIV would not want their identity to be leaked. Further, we show that the attack accuracy amplifies when the model is used to predict samples that come from a different distribution than the training set, which is often the case in real world applications. Therefore, we propose the use of *causal* learning approaches where a model learns the causal relationship between the input features and the outcome. Causal models are known to be invariant to the training distribution and hence generalize well to shifts between samples from the same distribution and across different distributions. First, we prove that models learned using causal structure provide stronger differential privacy guarantees than associational models under reasonable assumptions. Next, we show that causal models trained on sufficiently large samples are robust to membership inference attacks across different distributions of datasets and those trained on smaller sample sizes always have lower attack accuracy than corresponding associational models. Finally, we confirm our theoretical claims with experimental evaluation on 4 datasets with moderately complex Bayesian networks. We observe that neural network-based 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. Our results confirm the value of the generalizability of causal models in reducing susceptibility to privacy attacks.

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 particular 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, patients providing medical records to build a model that detects HIV 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 data is from the same distribution as the training dataset. 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 to detect HIV 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 privacy to different distributions and show that the risk from membership attack increases further on DNNs as the test distribution is changed. That is, the abiltity of an adversary to distinguish a member from a non-member improves with change in test distributions.

To alleviate privacy attacks, we propose using models that depend on the causal relationship between input features and the output. Causal learning has been extensively used to guarantee fairness and

explainability properties of the predicted output (Kusner et al., 2017; Nabi & Shpitser, 2018; Datta et al., 2016). However, the connection of causality to privacy 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 hence make the predictions of *causal models* independent of the observed data distribution, let alone the observed dataset. Hence, causal models generalize better even with change in the distributions.

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 a corresponding associational model trained on the same features and the same amount of added noise to the training dataset. Consequently, we show that membership inference attacks are ineffective (equivalent to a random guess) on causal models trained on infinite samples. Empirical attack accuracies on four different datasets confirm our theoretical claims. 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 increase as test distributions are changed. The attack accuracy reaches nearly 80% when the target associational model is trained on 60K training samples and used to predict test data that belong to a different distribution than the training data. Our results show that causal learning approaches are a promising direction for training models on sensitive data. Section 2 describes the properties of causal models. Section 3 proves the connections of causality to differential privacy and robustness to membership attacks. Section 4 provides empirical results. To summarize, we make the following contributions:

- For the same amount of added noise, models learned using causal structure provide stronger ϵ-differential privacy guarantees than corresponding associational models.
- Causal models are *provably* more robust to membership inference attacks than typical associational models such as neural networks.
- We simulate practical settings where the test distribution may not be the same as the training distribution and find that the membership inference attack accuracy of causal models is close to a "random guess" (i.e., 50%) while associational models exhibit upto 80% attack accuracy.

2 Properties of Causal Models

Causal models are shown to generalize well since the output of these models depend only on the causal relationship between the input features and the outcomes instead of the associations between them. From prior work, we know that the causal relationship between the features is invariant to the their distribution (Peters et al., 2016). Using this property, we study its effects on the privacy of 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; Bengio et al., 2019), 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 and $\mathtt{Y} \in \{0, 1\}$. 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. Figure 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, thus a generalizable model should not include these features. Similarly, Figure 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 $\mathtt{X}_{\mathtt{pa}}$. Therefore, including them in a model does not add predictive value (and further, avoids problems when the relationships between $\mathtt{X}_{\mathtt{S1}}$ and $\mathtt{X}_{\mathtt{S2}}$ may also change).

Finally, for completeness, the exhaustive set of variables to include is known as the *Markov Blanket*, $\mathtt{X_C}$ which includes $\mathtt{Y}$'s parents, ($\mathtt{X_{pa}}$), children ($\mathtt{Y_{ch}}$) and parents of children. Conditioned on its Markov blanket (Figure 1c), $\mathtt{Y}$ is independent of all other variables in the causal graph. When $\mathtt{Y}$ has no descendants in the graph, then the effective Markov blanket includes only its parents, $\mathtt{X_{pa}}$.

The key insight is that building a model for predicting Y using the Markov Blanket $\mathtt{X_C}$ 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 $\mathtt{X_C}$ to $\mathtt{Y}$ is stable. We call such a model as a *causal* model, the features in ($\mathtt{X_C}$) as the *causal features*, and assume that all the causal features for $\mathtt{Y}$ are observed. In contrast, we call a model that uses all available features as an *associational* model.

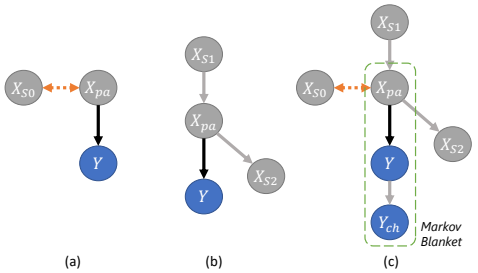


Figure 1: A causal predictive model includes only the parents of $\mathtt{Y}$ (a) and (b). Panel (c) shows the generalization to a Markov Blanket.

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, $\mathtt{L}(.,.)$ refers to the loss on a single input and $\mathcal{L}_{\mathtt{P}}(.,.) = \mathbb{E}_{\mathtt{P}}\mathtt{L}(.,.)$ refers to the expected value of the loss over a distribution $\mathtt{P}(\mathtt{X},\mathtt{Y})$. We refer $\mathtt{f} : \mathtt{X} \to \mathtt{Y}$ as the ground-truth labeling function and $\mathtt{h} : \mathtt{X} \to \mathtt{Y}$ as the hypothesis function or simply the model. Then, $\mathtt{L}(\mathtt{h},\mathtt{h}')$ is any loss function quantifying the difference between two models $\mathtt{h}$ and $\mathtt{h}'$.

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

$$\mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) = \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h},\mathtt{f}) \tag{1}$$

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

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

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

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

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

Theorem 1. *Consider a dataset* $\mathtt{S} \sim \mathtt{P}(\mathtt{X},\mathtt{Y})$, *a structural causal graph* $\mathtt{G}$ *that connects* $\mathtt{X}$ *to* $\mathtt{Y}$, *and causal features* $\mathtt{X_C}$ *where* $\mathtt{X_C}$ *is a Markov Blanket of* $\mathtt{Y}$ *under* $\mathtt{G}$. *Let* $\mathtt{P}^*(\mathtt{X},\mathtt{Y})$ *be a distribution with arbitrary* $\mathtt{P}^*(\mathtt{X})$ *such that the causal relationship from* $\mathtt{X_C}$ *to* $\mathtt{Y}$ *is preserved i.e.,* $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$. *Let* $\mathcal{H}_{\mathtt{C}}$ *represent the set of causal models* $\mathtt{h_c} : \mathtt{X_C} \to \mathtt{Y}$ *that use only causal features and* $\mathcal{H}_{\mathtt{A}}$ *represent the set of associational models* $\mathtt{h_a} : \mathtt{X} \to \mathtt{Y}$ *that use all the features, such that* $\mathcal{H}_C \subseteq \mathcal{H}_A$. *Then, for any symmetric loss function* $\mathtt{L}$ *that obeys the triangle inequality, the upper bound of* `ODE` *from* $\mathtt{S} \sim \mathtt{P}(\mathtt{X},\mathtt{Y})$ *to* $\mathtt{P}^*$ *(called* `ODE-Bound`*) for a causal model* $\mathtt{h_c} \in \mathcal{H}_{\mathtt{C}}$ *is less than or equal to the upper bound* `ODE-Bound` *of an associational model* $h_a \in \mathcal{H}_A$, *with probability at least* $(1-\delta)^2$.

$$\mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c},\mathtt{f};\delta) \leq \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a},\mathtt{f};\delta) \tag{4}$$

Proof. Since $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$, the optimal causal model that minimizes loss over $\mathtt{P}$ is the same as the loss-minimizing model over $\mathtt{P}^*$. That is, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$. However, for an associational model, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}$ and thus there is an additional loss term when generalizing to data from P^*. The rest of the proof follows from triangle inequality of the loss function and the standard bounds for `IDE` from past work. Detailed proof is in Appendix Section A.1. □

Corollary 1. *For a model* $\mathtt{h}$ *trained on a dataset* $\mathtt{S} \sim \mathtt{P}$*, and for any two input instances* $\mathtt{x} \sim \mathtt{P}$ *and* $\mathtt{x}' \sim \mathtt{P}^*$ *such that* $\mathtt{x} \in \mathtt{S}$ *, the worst-case generalization error for a causal model on any* $\mathtt{x}'$ *is less than or equal to that for an associational model.* [Proof in Appendix Section A.2]

$$\max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}' \sim \mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{f}) - \mathtt{L}_{\mathtt{x} \sim \mathtt{P}}(\mathtt{h}_\mathtt{c}, \mathtt{f}) \leq \max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}' \sim \mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{f}) - \mathtt{L}_{\mathtt{x} \sim \mathtt{P}}(\mathtt{h}_\mathtt{a}, \mathtt{f}) \tag{5}$$

3 Main Result: Privacy Guarantees 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 notion of privacy guarantees 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:* $\Pr(\mathcal{M}(d) \in \mathcal{S}) \leq e^\epsilon \Pr(\mathcal{M}(d') \in \mathcal{S})$

Based on the generalization property, we show that causal models provide stronger differential privacy guarantees than corresponding associational models. The standard approach to designing a differentially private algorithm is by calculating the sensitivity of that algorithm and adding noise proportional to the sensitivity. Sensitivity captures the change in the output of an algorithm due to the change in a single data point in the input. Higher the sensitivity, larger is the amount of noise required to make an algorithm differentially private with reasonable ϵ guarantees. We first provide the formal definition of sensitivity and then show that the sensitivity of causal models is lower than or equal to associational models.

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:* $\Delta\mathcal{F} = \max_{\mathtt{S},\mathtt{S}'} ||\mathcal{F}(\mathtt{S}) - \mathcal{F}(\mathtt{S}')||_1$

Lemma 1. *Given two neighboring datasets* $\mathtt{S}$ *and* $\mathtt{S}'$*, such that* $\mathtt{S}' = \mathtt{S} \backslash \mathtt{x} + \mathtt{x}'$ *where* $\mathtt{x} \in \mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$*,* $\mathtt{x}' \sim \mathtt{P}^*(\mathtt{X}, \mathtt{Y})$ *and* $\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{C})$*. Let a model* 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. Then, the sensitivity of a causal learning function* $\mathcal{F}_c$ *that outputs learnt empirical hypothesis* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} \leftarrow \mathcal{F}_\mathtt{c}(\mathtt{S})$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'} \leftarrow \mathcal{F}_\mathtt{c}(\mathtt{S}')$ *is lower than or equal to the sensitivity of an associational learning function* $\mathcal{F}_a$ *that outputs* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}} \leftarrow \mathcal{F}_\mathtt{a}(\mathtt{S})$ *and* $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'} \leftarrow \mathcal{F}_\mathtt{a}(\mathtt{S}')$ *using a loss function* $\mathtt{L}$ *that is strongly convex over* Ω*, symmetric and obeys the triangle inequality.*

$$\Delta\mathcal{F}_\mathtt{c} = \max_{\mathtt{S},\mathtt{S}'} ||\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}||_1 \leq \max_{\mathtt{S},\mathtt{S}'} ||\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}||_1 = \Delta\mathcal{F}_\mathtt{a} \tag{6}$$

Proof. We can write the empirical loss minimizers for the datasets $\mathtt{S}$ and $\mathtt{S}'$ as:

$$\begin{aligned} \mathtt{h}^{\mathtt{min}}_\mathtt{S} &= \arg\min_\mathtt{h} \mathcal{L}_\mathtt{S}(\mathtt{h}, \mathtt{f}) = \arg\min_\mathtt{h} \frac{1}{\mathtt{N}} \sum_{\mathtt{i}=1}^{\mathtt{N}} \mathtt{L}_{\mathtt{x}_\mathtt{i}}(\mathtt{h}, \mathtt{f}) \\ \mathtt{h}^{\mathtt{min}}_{\mathtt{S}'} &= \arg\min_\mathtt{h} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}, \mathtt{f}) = \arg\min_\mathtt{h} \frac{1}{\mathtt{N}} \sum_{\mathtt{i}=1}^{\mathtt{N}-1} \mathtt{L}_{\mathtt{x}_\mathtt{i}}(\mathtt{h}, \mathtt{f}) + [\mathtt{L}_{\mathtt{x}'}(\mathtt{h}, \mathtt{f}) - \mathtt{L}_\mathtt{x}(\mathtt{h}, \mathtt{f})] \end{aligned} \tag{7}$$

From Corollary 1, for $\mathtt{x}' \sim \mathtt{P}^*$ and $\mathtt{x} \sim \mathtt{P}$, we have:

$$\max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}' \sim \mathtt{P}^*}(\mathtt{h}_\mathtt{c}, \mathtt{f}) - \mathtt{L}_{\mathtt{x} \sim \mathtt{P}}(\mathtt{h}_\mathtt{c}, \mathtt{f}) \leq \max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}' \sim \mathtt{P}^*}(\mathtt{h}_\mathtt{a}, \mathtt{f}) - \mathtt{L}_{\mathtt{x} \sim \mathtt{P}}(\mathtt{h}_\mathtt{a}, \mathtt{f}) \tag{8}$$

Since $\mathtt{x} \in \mathtt{S}$ and $\mathtt{x}' \in \mathtt{S}'$ and $|\mathtt{S} - \mathtt{S}'| = 1$ the above is true for any $\mathtt{S}$ and $\mathtt{S}'$,

$$\max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}_\mathtt{c}, \mathtt{f}) - \mathcal{L}_\mathtt{S}(\mathtt{h}_\mathtt{c}, \mathtt{f}) \leq \max_{\mathtt{S},\mathtt{S}'} \mathcal{L}_{\mathtt{S}'}(\mathtt{h}_\mathtt{a}, \mathtt{f}) - \mathcal{L}_\mathtt{S}(\mathtt{h}_\mathtt{a}, \mathtt{f}) \tag{9}$$

Since $\mathtt{L}$ is a strongly convex function over Ω, and since $\mathcal{H}_C \subseteq \mathcal{H}_A \Rightarrow \Omega_C \subseteq \Omega_A$, $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}$ that minimize $\mathcal{L}_\mathtt{S}(\mathtt{h}, \mathtt{f})$ and $\mathcal{L}_{\mathtt{S}'}(\mathtt{h}, \mathtt{f})$ respectively should also be closer to each other than $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}}$ and $\mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}$ (Boyd & Vandenberghe, 2004) (using Eqn. 9).

$$\max_{\mathtt{S},\mathtt{S}'} || \theta^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} - \theta^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}||_1 \leq \max_{\mathtt{S},\mathtt{S}'} || \theta^{\mathtt{min}}_{\mathtt{a},\mathtt{S}} - \theta^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}||_1 \Rightarrow \max_{\mathtt{S},\mathtt{S}'} || \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{c},\mathtt{S}'}||_1 \leq \max_{\mathtt{S},\mathtt{S}'} || \mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}} - \mathtt{h}^{\mathtt{min}}_{\mathtt{a},\mathtt{S}'}||_1 \tag{10}$$

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

Theorem 2. *Let $\hat{\mathcal{F}}_c$ and $\hat{\mathcal{F}}_a$ be the differentially private algorithms corresponding to causal learning and associational learning algorithms $\mathcal{F}_c$ 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 noise sampled from the same distribution,* $\mathtt{Lap(Z)}$ *for both algorithms, 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,

$$\hat{\mathcal{F}}_c = \mathcal{F}_{\mathtt{c}} + \mathcal{K} \sim \mathtt{Lap}(\frac{\Delta\mathcal{F}_{\mathtt{c}}}{\epsilon_{\mathtt{c}}}) \qquad \hat{\mathcal{F}}_a = \mathcal{F}_{\mathtt{a}} + \mathcal{K} \sim \mathtt{Lap}(\frac{\Delta\mathcal{F}_{\mathtt{a}}}{\epsilon_{\mathtt{a}}}) \tag{11}$$

Since $\mathcal{K}$ is sampled from the same noise distribution,

$$\mathtt{Lap}(\frac{\Delta\mathcal{F}_{\mathtt{c}}}{\epsilon_{\mathtt{c}}}) = \mathtt{Lap}(\frac{\Delta\mathcal{F}_{\mathtt{a}}}{\epsilon_{\mathtt{a}}}) \qquad \therefore \frac{\Delta\mathcal{F}_{\mathtt{c}}}{\epsilon_{\mathtt{c}}} = \frac{\Delta\mathcal{F}_{\mathtt{a}}}{\epsilon_{\mathtt{a}}} \tag{12}$$

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 true for models developed using recent work (Papernot et al., 2017) that provides a tighter data-dependent differential privacy guarantee. The key idea is to produce an output label based on voting from M teacher models, each trained on a disjoint subset of the training data. We state the theorem below and provide the proof in Appendix B. Given datasets from different domains, the below theorem provides a constructive proof to generate a differentially private causal algorithm, following the method from Papernot et al. (2017).

Theorem 3. *Let* $\mathtt{D}$ *be a dataset generated from possibly a mixture of different distributions* $\mathtt{P(X,Y)}$ *such that* $\mathtt{P(Y|X_C)}$ *remains the same. Let n_j be the votes for the jth class from M teacher models. Let $\mathcal{M}$ be the mechanism that produces a noisy max,* $\arg\max_j\{n_j + Lap(2/\gamma)\}$. *Then the privacy budget ϵ for a causal model is lower than that for the associational model with the same accuracy.*

3.2 Robustness to Membership Attacks

Deep learning models have 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 access to a black-box model only sees the confidence scores for the predicted output whereas one with the white-box has access to the model parameters and observe 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). For the white-box setting, if an adversary knows the true label for the target input, then they may guess the input to be a member of the training set whenever the loss is lower, and vice-versa. 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 for any new input (Salem et al., 2018).

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 to detect HIV, the test inputs may come from different hospitals. 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. From Yeom et al. (2018), we first define a membership attack as:

Definition 6. *Let model h be trained on a dataset* $\mathtt{S(X,Y)}$ *of size N. Let $\mathcal{A}$ be an adversary with access to h and a input X. The 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.*

$$\mathtt{Adv}(\mathcal{A},\mathtt{h}) = \Pr[\mathcal{A} = \mathtt{1}|\mathtt{b} = \mathtt{1}] - \Pr[\mathcal{A} = \mathtt{1}|\mathtt{b} = \mathtt{0}] \tag{13}$$

where $b = 1$ if the input is in the training set and 0 otherwise.

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

Theorem 4. *Given a structural causal network that connects* $\mathtt{X}$ *to* $\mathtt{Y}$*, let* $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$ *be a dataset sampled from* $\mathtt{P}$*, and let* $\mathtt{P}^*$ *be any distribution such that* $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$*. Then, a causal model* $\mathtt{h_c}$ *trained on* $\mathtt{S}$ *yields lower membership advantage than an associational model* h_a *even when the test dataset is from a different distribution* $\mathtt{P}^*$.

Proof. From Theorem 2 above, we can construct an ϵ_c-DP mechanism based on a causal model, and a ϵ_a-DP mechanism based on an associational model, where $\epsilon_c \leq \epsilon_a$. Further, this construction works for different input distributions. From Lemma 2, the membership advantage of an adversary $\mathcal{A}$ is,

$$\mathtt{Adv}(\mathcal{A}, \mathtt{h_c}) \leq \mathtt{exp}(\epsilon_\mathtt{c}) - 1 \qquad \mathtt{Adv}(\mathcal{A}, \mathtt{h_a}) \leq \mathtt{exp}(\epsilon_\mathtt{a}) - 1 \tag{14}$$

Thus, worst case advantage for a causal model is always lower than that of an associational model. □

Corollary 2. *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* $\mathtt{Adv}(\mathcal{A}, \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) \to 0$.

The proof is the 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 $N \to \infty$ 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 Appendix Section C.

Attribute Inference attacks. We prove similar results on the benefits of causal models for attribute inference attacks in Appendix Section D.

4 Implementation and Evaluation

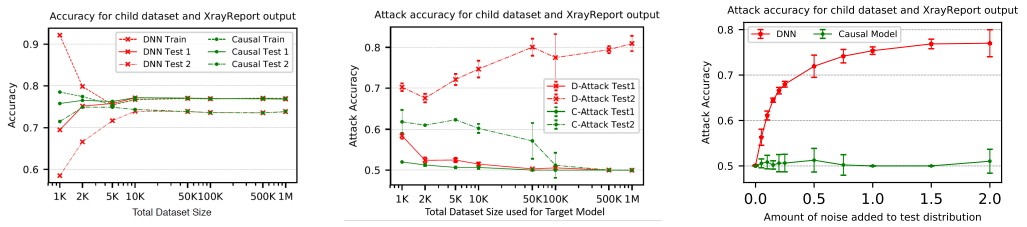
Benchmark Datasets. To avoid errors in learning causal structure from data, we perform evaluation on datasets for which the causal structure and the true conditional probabilities of the variables are known from prior research. We select 4 Bayesian network datasets— Child, Sachs, Alarm and Water that range from 230-10k parameters (Table 1) (bnl). 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 all 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.

Table 1: Details of the benchmark datasets

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

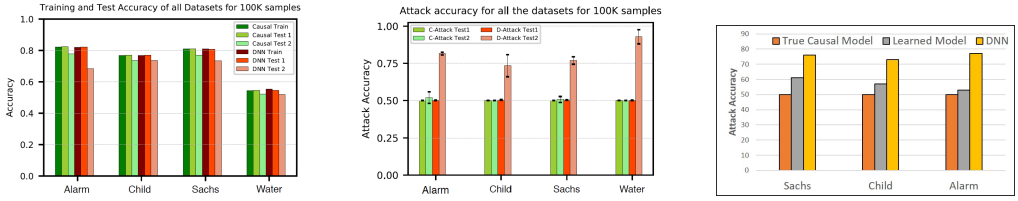
Implementation. We sample data using the causal structure and probabilities from the Bayesian network, and use a 60:40% split for train-test datasets. We learn a causal model and a deep neural network (DNN) on each training dataset. We implement the attacker model to perform membership inference attack 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 4 in Appendix. Note that the attack accuracies reported are an upper bound since we assume that the adversary has white-box access to 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 ten. 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.



(a) (b) (c)

Figure 2: 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 3: (a)Target accuracy, (b) Attack accuracy, (c) Attack accuracy for true, learned causal model and DNN.

4.1 Experimental Setup

We evaluate the DNN and the causal model sample sizes ranging from 1K to 1M dataset sizes. We refer Test 1 as the test dataset which is drawn from the same distribution as the training data and Test 2 is generated from a completely different distribution except for the relationship of the output class to its parents. To generate Test 2, we alter the true probabilities $\Pr(X)$ uniformly at random (later, we consider adding noise to the original value). Our goal with generating Test 2 is to capture the realistic behaviour of shift in distribution for input features. We refer the causal and DNN model as the *target* on which the attack is perpetrated.

4.2 Results

We present results on the accuracy of target models (causal and DNN models) and the membership attack accuracy for different dataset sizes and test distributions.

Accuracy comparison of DNN and Causal models. Figure 2a 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 1 and Test 2) 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 1 dataset have the same accuracy. The accuracy for the Test 2 distribution stabilizes for a total dataset size of 10K samples. In contrast, for the causal model, the train and Test 1 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 2 dataset is the same for both the DNN and the causal model. Figure 3a shows the accuracy comparison for all four datasets with similar results.

Attack Accuracy comparison 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 2b shows the attack accuracy comparison for Test 1 (same distribution) and Test 2 (different distribution) datasets. Attack accuracy of the Test 1 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 1 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 2, 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 (as shown in Figure 2a). These results show that causal models generalize better than DNNs across input distributions. Figure 3b 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 2 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 also 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 probabilities which are then normalized to sum up to 1. Figure 2c shows the comparison of 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 be different from the true causal structure. We evaluate the attack accuracy for learned causal models on the `Sachs, Child` and `Alarm` dataset[1]. 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 3c. 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 in Appendix gives a fine-grained analysis.

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 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, as shown in Section 4, causal models are, by definition, robust to membership inference attacks across distributions.

Causal learning and privacy. There is a 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 applies causal models to study properties of machine learning models such as providing explanations (Datta et al., 2016) or fairness (Kusner et al., 2017), the relation of causality to privacy is yet unexplored. With this paper, we present the first result which shows the privacy benefits of causal models.

6 Conclusion and Future Work

We conclude that causal learning is a promising approach to train models which are robust to privacy attacks such as membership inference and model inversion. As our future work, we want to relax our assumption of a known causal structure and investigate the privacy guarantees of causal models where the causal features and the relationship between them is not known apriori (Peters et al., 2017).

[1]We exclude the `Water` dataset as `bn.fit` in bnlearn library gives error due to the extreme probabilities.

REFERENCES

Bayesian Network Repository. www.bnlearn.com/bnrepository/.

Tensorflow. https://www.tensorflow.org/.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232*, 2018.

Anupam Datta, Shayak Sen, and Yair Zick. 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.

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

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

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

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 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.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pp. 555–563, 2016.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Matt J Kusner, Yu Sun, Karthik Sridharan, and Kilian Q Weinberger. Private causal inference. *arXiv preprint arXiv:1512.05469*, 2015.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Razieh Nabi and Ilya Shpitser. 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.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 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.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 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.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.

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

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

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 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.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Md Atiqur Rahman, Tanzila Rahman, Robert Laganiere, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 2018.

Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

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

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 3–18. IEEE, 2017.

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

Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Using deep learning to detect price change indications in financial markets. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2511–2515. IEEE, 2017.

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

A Generalization Properties of Causal Model

A.1 Generalization over Different Distributions

Theorem 1. *Consider a dataset* $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$*, a structural causal graph* $\mathtt{G}$ *that connects* $\mathtt{X}$ *to* $\mathtt{Y}$*, and causal features* $\mathtt{X_C}$ *where* $\mathtt{X_C}$ *is a Markov Blanket of* $\mathtt{Y}$ *under* $\mathtt{G}$*. Let* $\mathtt{P}^*(\mathtt{X}, \mathtt{Y})$ *be a distribution with arbitrary* $\mathtt{P}^*(\mathtt{X})$ *such that the causal relationship from* $\mathtt{X_C}$ *to* $\mathtt{Y}$ *is preserved i.e.,* $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$*. Let* $\mathcal{H}_\mathtt{C}$ *represent the set of causal models* $\mathtt{h_c} : \mathtt{X_C} \to \mathtt{Y}$ *that use only causal features and* $\mathcal{H}_\mathtt{A}$ *represent the set of associational models* $\mathtt{h_a} : \mathtt{X} \to \mathtt{Y}$ *that use all the features, such that* $\mathcal{H}_C \subseteq \mathcal{H}_A$*. Then, for any symmetric loss function* $\mathtt{L}$ *that obeys the triangle inequality, the upper bound of* $\mathtt{ODE}$ *from* $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$ *to* $\mathtt{P}^*$ *(called* $\mathtt{ODE\text{-}Bound}$*) for a causal model* $\mathtt{h_c} \in \mathcal{H}_\mathtt{C}$ *is less than or equal to the upper bound* $\mathtt{ODE\text{-}Bound}$ *of an associational model* $h_a \in \mathcal{H}_A$*, with probability at least* $(1-\delta)^2$*.*

$$\mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_c}, \mathtt{f}; \delta) \leq \mathtt{ODE\text{-}Bound}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h_a}, \mathtt{f}; \delta) \tag{4}$$

Proof. For any function $\mathtt{h} : \mathtt{X} \to \mathtt{Y}$ that was trained on $\mathtt{S} \sim \mathtt{P}(\mathtt{X}, \mathtt{Y})$, from Def. 2 we write:

$$\begin{aligned}
\mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h}, \mathtt{f}) &= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}, \mathtt{f}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h}, \mathtt{f}) \\
&= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}, \mathtt{f}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h}, \mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}, \mathtt{f}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h}, \mathtt{f}) \\
&= \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}, \mathtt{f}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h}, \mathtt{f}) + \mathtt{IDE}_{\mathtt{P}}(\mathtt{h}, \mathtt{f})
\end{aligned} \tag{15}$$

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 $\mathtt{P}$ and $\mathtt{P}^*$ as $\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}$ and $\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}}$.

$$\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}\in\mathcal{H}} \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) \qquad \mathtt{h}_{\mathtt{P}_*}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}\in\mathcal{H}} \mathcal{L}_{\mathtt{P}*}(\mathtt{h},\mathtt{f}) \tag{16}$$

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

$$\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{f}) \leq \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}},\mathtt{f}) \tag{17}$$

And,

$$\begin{aligned} \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) &\geq \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f}) \\ \Rightarrow -\mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) &\leq -\mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f}) \end{aligned} \tag{18}$$

Thus, combining Eq. 15, 17 and 18, we obtain,

$$\begin{aligned} \mathtt{ODE}_{\mathtt{P},\mathtt{P}^*}(\mathtt{h},\mathtt{f}) &\leq \mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}},\mathtt{f}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f}) \\ &= \mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) + (\mathcal{L}_{\mathtt{P}^*}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}}) - \mathcal{L}_{\mathtt{P}}(\mathtt{h},\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}})) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}},\mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}}) \\ &\leq \mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{f}) + \mathtt{disc}_{\mathtt{L},\mathcal{H}}(\mathtt{P},\mathtt{P}^*) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}},\mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f}) + \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}}) \end{aligned} \tag{19}$$

where the last inequality is due to the definition of discrepancy distance (Definition 3). Equation 19 divides the out-of-distribution generalization error of a hypothesis $\mathtt{h}$ in four parts:

1. $\mathtt{IDE}_{\mathtt{P}}(\mathtt{h},\mathtt{f})$ 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 $\mathcal{S}$ (Shalev-Shwartz & Ben-David, 2014). Using a uniform error bound based on the VC dimension, we obtain, with probability at least $1-\delta$,

$$\mathtt{IDE} \leq \sqrt{8\frac{\mathtt{VCdim}(\mathcal{H})(\ln(2|\mathcal{S}|)+1)+\ln(4/\delta)}{|\mathcal{S}|}} = \mathtt{IDE\text{-}Bound}(\mathcal{H},\mathcal{S}) \tag{20}$$

   Since $H_C \subseteq H_A$, VC-dimension of causal models is not greater than that of associational models. Thus,

$$\mathtt{VCDim}(\mathcal{H}_{\mathtt{C}}) \leq \mathtt{VCDim}(\mathcal{H}_{\mathtt{A}}) \Rightarrow \mathtt{IDE\text{-}Bound}(\mathcal{H}_{\mathtt{C}},\mathcal{S}) \leq \mathtt{IDE\text{-}Bound}(\mathcal{H}_{\mathtt{A}},\mathcal{S}) \tag{21}$$

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 hypothesis class $\mathcal{H}$.
3. $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}},\mathtt{f})$ and $\mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{f})$ measure the error due to the true labeling function $\mathtt{f}$ being outside the hypothesis class $\mathcal{H}$.
4. $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}})$ denotes the loss (or difference) between the loss-minimizing function over $\mathtt{P}$ and the loss-minimizing hypothesis over $\mathtt{P}^*$.

We next show that for a given distribution $\mathtt{P}$ and another distribution $\mathtt{P}^*$ such that $\mathtt{P}(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_{\mathtt{C}})$ where $\mathtt{X}_{\mathtt{C}}$ is the Markov Blanket for $\mathtt{Y}$, the last term $\mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{P}^*}^{\mathtt{OPT}})$ of Equation 19 vanishes for a causal model but may remain non-zero for an associational model.

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

$$\Pr(\mathtt{Y}|\mathtt{do}(\mathtt{X}_{\mathtt{c}}=\mathtt{x}_{\mathtt{c}})) = \mathtt{P}(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}=\mathtt{x}_{\mathtt{c}}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_{\mathtt{C}}=\mathtt{x}_{\mathtt{c}}) \tag{22}$$

where the last equality is since $\mathtt{X}_{\mathtt{C}}$ is also the Markov Blanket in $\mathtt{P}^*$, ($\mathtt{Y}$ is independent of $\mathtt{X}_{\mathcal{C}'}$ given $\mathtt{X}_{\mathtt{C}}$ in $\mathtt{P}^*$), and thus $\mathtt{X}_{\mathtt{C}}$ includes all parents of $\mathtt{Y}$. Since the conditional distribution of Y given X_C features is the same across P and P^*, the loss-minimizing models should also be the same. Defining $\mathtt{h}_{\mathtt{c},\mathtt{P}}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}_{\mathtt{c}}\in\mathcal{H}_{\mathtt{C}}} \mathcal{L}_{\mathtt{P}}(\mathtt{h}_{\mathtt{c}},\mathtt{f})$ and $\mathtt{h}_{\mathtt{c},\mathtt{P}^*}^{\mathtt{OPT}} = \arg\min_{\mathtt{h}_{\mathtt{c}}\in\mathcal{H}_{\mathcal{C}}} \mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{c}},\mathtt{f})$, we obtain,

$$\mathtt{h}_{\mathtt{c},\mathtt{P}}^{\mathtt{OPT}} = \mathtt{h}_{\mathtt{c},\mathtt{P}^*}^{\mathtt{OPT}} \tag{23}$$

And therefore, the term $\mathcal{L}_{\mathtt{P}^*}(\mathtt{h}_{\mathtt{c},\mathtt{P}}^{\mathtt{OPT}},\mathtt{h}_{\mathtt{c},\mathtt{P}^*}^{\mathtt{OPT}}) = 0$.

Associational Model. In contrast, an associational model may use a subset of $\mathtt{X}$, $\mathtt{X_S} \subseteq \mathtt{X}$, that may not include all variables in the Markov Blanket, or may include the Markov Blanket but also include other extraneous variables. When at least one parent is excluded, the subset no longer corresponds to the causal do distribution.

$$\mathtt{P}(\mathtt{Y}|\mathtt{X_S} = \mathtt{x_s}) \neq \Pr[\mathtt{Y}|\mathtt{do}(\mathtt{X_S} = \mathtt{x_s})] \neq \mathtt{P}^*(\mathtt{Y}|\mathtt{X_S} = \mathtt{x_s}) \tag{24}$$

In such a case, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}$. Since $\mathtt{X_C} \subseteq \mathtt{X}$, it is possible in some cases that an associational model includes all the causal features of $\mathtt{Y}$. Therefore, $\mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}) \geq 0$.

Loss Bounds. Further, since $\mathtt{X_C}$ is the Markov Blanket and thus the set of strongly relevant features as well (Pellet & Elisseeff, 2008). From the literature we know that, $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}(\mathtt{Y}|\mathtt{X_C}, \mathtt{X})$, hence,

$$\mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{f}) \leq \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{f}) \tag{25}$$

And similarly,

$$\mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P*}}, \mathtt{f}) \leq \mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}, \mathtt{f}) \tag{26}$$

Hence, using Eq. 23, 25, 26, we write equation 19 for a causal model as:

$$\begin{aligned}
\mathtt{ODE}_{\mathtt{P,P^*}}(\mathtt{h_c}, \mathtt{f}) &= \mathcal{L}_{\mathtt{P^*}}(\mathtt{h_c}, \mathtt{f}) - \mathcal{L}_{\mathtt{S}\sim\mathtt{P}}(\mathtt{h_c}, \mathtt{f}) \\
&\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{c}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE}_{\mathtt{P}}(\mathtt{h_c}, \mathtt{f}) + \mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P^*}}, \mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}}, \mathtt{f}) \\
&\leq \mathtt{disc}_{\mathtt{L},\mathcal{H}_\mathtt{c}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE}_{\mathtt{P}}(\mathtt{h_c}, \mathtt{f}) + \mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}, \mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{f})
\end{aligned} \tag{27}$$

Using Eqns. 20 and 21, we obtain, with probability at least $1 - \delta$:

$$\begin{aligned}
\mathtt{ODE}_{\mathtt{P,P^*}}(\mathtt{h_c}, \mathtt{f}) &\leq \mathtt{disc}_{\mathtt{L},\mathcal{H_C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_{\mathtt{P}}(\mathcal{H_C}, \mathcal{S}; \delta) + \mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}, \mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{f}) && (28) \\
&\leq \mathtt{disc}_{\mathtt{L},\mathcal{H_C}}(\mathtt{P}, \mathtt{P}^*) + \mathtt{IDE{-}Bound}_{\mathtt{P}}(\mathcal{H_A}, \mathcal{S}; \delta) + \mathcal{L}_{\mathtt{P^*}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}, \mathtt{f}) + \mathcal{L}_{\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}}, \mathtt{f}) && (29) \\
&= \mathtt{ODE{-}Bound}_{\mathtt{P,P^*}}(\mathtt{h_c}, \mathtt{f}; \delta) && (30)
\end{aligned}$$

Similarly, for an associational model, we obtain, with probability at least $1 - \delta$:

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

Finally, using Definition 3 for discrepancy distance, we can state $\mathcal{H_C} \subseteq \mathcal{H_A} \Rightarrow \mathtt{disc}_{\mathtt{L},\mathcal{H_C}}(\mathtt{P}, \mathtt{P}^*) \leq \mathtt{disc}_{\mathtt{L},\mathcal{H_A}}(\mathtt{P}, \mathtt{P}^*)$. Therefore, from Eq. 27 and 31, we claim with probability $(1 - \delta)^2$,

$$\mathtt{ODE{-}Bound}_{\mathtt{P,P^*}}(\mathtt{h_c}, \mathtt{f}; \delta) \leq \mathtt{ODE{-}Bound}_{\mathtt{P,P^*}}(\mathtt{h_a}, \mathtt{f}; \delta) \tag{32}$$

□

A.2 Generalization over a Single Datapoint

Corollary 1. *For a model* $\mathtt{h}$ *trained on a dataset* $\mathtt{S} \sim \mathtt{P}$*, and for any two input instances* $\mathtt{x} \sim \mathtt{P}$ *and* $\mathtt{x}' \sim \mathtt{P}^*$ *such that* $\mathtt{x} \in \mathtt{S}$ *, the worst-case generalization error for a causal model on any* $\mathtt{x}'$ *is less than or equal to that for an associational model.* [Proof in Appendix Section A.2]

$$\max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'\sim\mathtt{P^*}}(\mathtt{h_c}, \mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h_c}, \mathtt{f}) \leq \max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'\sim\mathtt{P^*}}(\mathtt{h_a}, \mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h_a}, \mathtt{f}) \tag{5}$$

Proof. Using the triangle inequality for the loss function, we obtain,

$$\mathtt{L}_{\mathtt{x}'\sim\mathtt{P^*}}(\mathtt{h}, \mathtt{f}) \leq \mathtt{L}_{\mathtt{x}'}(\mathtt{h}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{P^*}}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P^*}}, \mathtt{f}) \tag{33}$$

$$-\mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h}, \mathtt{f}) \leq -\mathtt{L}_{\mathtt{x}}(\mathtt{h}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}, \mathtt{f}) \tag{34}$$

Combining the two inequalities,

$$\begin{aligned}\mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h},\mathtt{f}) &\leq \mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*},\mathtt{f}) \\ &\quad - \mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{f}) \\ &\leq \mathtt{dist}_{\mathcal{H}}(\mathtt{x},\mathtt{x}') + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{f}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*})\end{aligned} \tag{35}$$

where $\mathtt{dist}(\mathtt{x},\mathtt{x}') = \max_{\mathtt{h},\mathtt{h}'\in\mathcal{H}} |\mathtt{L}_{\mathtt{x}'}(\mathtt{h},\mathtt{h}') - \mathtt{L}_{\mathtt{x}}(\mathtt{h},\mathtt{h}')|$, analogous to the $\mathtt{disc}_{\mathtt{L}}$ for distributions.

For a causal model, we know that $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}$. In addition, since $\mathtt{X}_{\mathtt{C}}$ is also the strongly relevant feature set (Pellet & Elisseeff, 2008), $\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}},\mathtt{f}) \leq \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f})$ and same for $\mathtt{P}^*$ and loss on $\mathtt{x}'$.

Hence,

$$\mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{p}}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) \leq \mathtt{dist}_{\mathcal{H}_{\mathtt{C}}}(\mathtt{x},\mathtt{x}') + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) \tag{36}$$

$$\mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}_{\mathtt{a}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{p}}(\mathtt{h}_{\mathtt{a}},\mathtt{f}) \leq \mathtt{dist}_{\mathcal{H}_{\mathtt{A}}}(\mathtt{x},\mathtt{x}') + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}) \tag{37}$$

where $\mathtt{dist}_{\mathcal{H}_{\mathtt{C}}}(\mathtt{x},\mathtt{x}') \leq \mathtt{dist}_{\mathcal{H}_{\mathtt{A}}}(\mathtt{x},\mathtt{x}')$ since $\mathcal{H}_{\mathtt{C}} \subseteq \mathcal{H}_{\mathtt{A}}$.

Next, we show that these bounds are tight. That is, there exists an associational model $\mathtt{h}_{\mathtt{a}}$ whose generalization error on $\mathtt{x}'$ is exactly the RHS on Eqn. 37 and thus higher than the bound for any causal model. Below we prove by constructing one such $\mathtt{h}_{\mathtt{a}}$.

For simplicity in construction, let us select $\mathtt{H}_{\mathtt{C}} = \{\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}\}$ and $\mathtt{H}_{\mathtt{A}} = \{\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c},\mathtt{P}^*}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}}, \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}\}$. Thus $\mathtt{dist}_{\mathtt{H}_{\mathtt{C}}}(\mathtt{x},\mathtt{x}') = 0$ whereas $\mathtt{dist}_{\mathtt{H}_{\mathtt{A}}}(\mathtt{x},\mathtt{x}') \geq 0$. We obtain the ODE-generalization bound for a causal model,

$$\mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{p}}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) \leq \mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) = \mathtt{ODE\text{-}Bound}_{\mathtt{x},\mathtt{x}'}(\mathtt{h}_{\mathtt{c}},\mathtt{f};\mathtt{H}_{\mathtt{C}}) \tag{38}$$

Let us now construct an associational model, $\mathtt{h}^{\dagger}_{\mathtt{a}}$, such that:

$$\begin{aligned}\mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{f}) &= \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*}) + \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}^*},\mathtt{f}) \\ -\mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{f}) &= -\mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}},\mathtt{f})\end{aligned} \tag{39}$$

A simple construction for the above equalities is to select $\mathtt{P}$ and $\mathtt{P}^*$, and correspondingly $\mathtt{x}^{\dagger}$ and $\mathtt{x}'^{\dagger}$ such that $\mathtt{L}_{\mathtt{x}^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) = \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) = 0$ and $\mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}) > 0$. Further, $\mathtt{h}^{\dagger}_{\mathtt{a}}$ can be selected such that $\mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) = 0$ and $\mathtt{L}_{\mathtt{x}^{\dagger}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) = 0$.

Then, using Eqn. 39,

$$\begin{aligned}\mathtt{L}_{\mathtt{x}'^{\dagger}\sim\mathtt{P}^*}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}^{\dagger}\sim\mathtt{p}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{f}) &= \mathtt{L}_{\mathtt{x}'}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) - \mathtt{L}_{\mathtt{x}}(\mathtt{h}^{\dagger}_{\mathtt{a}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{P}}) + \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) + \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}) \\ &= 0 + \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) + \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*}) \\ &> \mathtt{L}_{\mathtt{x}'^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}^*},\mathtt{f}) + \mathtt{L}_{\mathtt{x}^{\dagger}}(\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a},\mathtt{P}},\mathtt{f}) \\ &= \mathtt{ODE\text{-}Bound}_{\mathtt{x},\mathtt{x}'}(\mathtt{h}_{\mathtt{c}},\mathtt{f};\mathtt{H}_{\mathtt{C}})\end{aligned} \tag{40}$$

where the last equality comes from Eqn. 38. Combining Eqns. 40 and 36, we obtain,

$$\max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h}_{\mathtt{c}},\mathtt{f}) \leq \max_{\mathtt{x},\mathtt{x}'} \mathtt{ODE\text{-}Bound}_{\mathtt{x},\mathtt{x}'}(\mathtt{h}_{\mathtt{c}},\mathtt{f};\mathtt{H}_{\mathtt{C}}) \leq \max_{\mathtt{x},\mathtt{x}'} \mathtt{L}_{\mathtt{x}'\sim\mathtt{P}^*}(\mathtt{h}_{\mathtt{a}},\mathtt{f}) - \mathtt{L}_{\mathtt{x}\sim\mathtt{P}}(\mathtt{h}_{\mathtt{a}},\mathtt{f}) \tag{41}$$

□

B 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. We first present a Lemma on generalization of causal models and then prove the main result.

As a consequence of Theorem 1, another generalization property of causal models is that the models trained on data from two different distributions $P(X)$ and $P^*(X)$ are likely to output the same value for a new input.

Lemma 3. *Under the conditions of Theorem 1, let $h_{c,P}$ be a causal model trained on distribution P and let h_{c,P^*} be a model trained on P^*. Similarly, let $h_{a,P}$ and h_{a,P^*} be correlational models trained on P and P^* respectively. Assume that the correlational and causal models report the same accuracy α. Then for any new data input x,*

$$\Pr(h_{c,P}(x) = h_{c,P^*}(x)) \geq \Pr(h_{a,P}(x) = h_{a,P^*}(x))$$

As the size of the training sample $N \to \infty$, the LHS$\to 1$.

Proof. Let $h_{a,P}^{OPT} = \arg\min_{h \in \mathcal{H}_A} \mathcal{L}_P(h, f)$ and similarly let $h_{a,P^*}^{OPT} = \arg\min_{h \in \mathcal{H}_A} \mathcal{L}_{P^*}(h, f)$ be the loss-minimizing hypotheses under these two distributions, where $\mathcal{H}$ is the set of hypotheses. We can analogously define $h_{c,P}^{OPT}$ and h_{c,P^*}^{OPT}. For a causal model, we know from Theorem 1 that $h_{c,P}^{OPT} = h_{c,P^*}^{OPT}$. As $N \to \infty$, each of models on P and P^* approach their loss-minimizing functions. Then, for any input x,

$$L(h_{c,P}(x), h_{c,P^*}(x)) = L(h_{c,P}^{OPT}(x), h_{c,P^*}^{OPT}(x)) = 0 \tag{42}$$

$$L(h_{a,P}(x), h_{a,P^*}(x)) = L(h_{a,P}^{OPT}(x), h_{a,P^*}^{OPT}(x)) \geq 0 \tag{43}$$

$$\Rightarrow \Pr(h_{c,P}(x) = h_{c,P^*}(x)) = 1 \geq \Pr(h_{a,P}(x) = h_{a,P^*}(x)) \tag{44}$$

Under finite samples, let the accuracy of each model be α. Then for any new input, the causal models predict the same output if they both match $h_{c,P}^{min} = h_{c,P^*}^{min}$ or both do not match. The probability of this event can be captured by the accuracy of the models: α^2 if both models match $h_{c,P}^{min}$, and $(1-\alpha)^2$ if they do not.

$$\Pr(h_{c,P} = h_{c,P^*}) = \alpha^2 + (1-\alpha)^2 \tag{45}$$

For correlational models, $h_{a,P}^{OPT} \neq h_{a,P^*}^{OPT}$ for some x. For those x, assuming the same accuracy α,

$$\Pr(h1 = h2) = 2\alpha(1-\alpha) \tag{46}$$

Since $2\alpha(1-\alpha) \leq \alpha^2 + (1-\alpha)^2 \; \forall \alpha \in [0, 1]$, we have

$$\Pr(h_{c,P} = h_{c,P^*}) \geq \Pr(h_{a,P} = h_{a,P^*})$$

□

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 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 M models. The privacy guarantees are better if a larger number of teacher models agree on each input, since by definition the majority decision could not have been changed by modifying a single data point (or a single teacher's vote). Since causal models generalize to new distributions, intuitively we expect causal models trained on separate samples to agree more. Below we show that for a fixed amount of noise, a causal model is ϵ_c-DP compared to ϵ-DP for a associational model, where $\epsilon_c \leq \epsilon$.

Theorem 3. *Let $\mathtt{D}$ be a dataset generated from possibly a mixture of different distributions $\mathtt{P(X, Y)}$ such that $\mathtt{P(Y|X_C)}$ remains the same. Let n_j be the votes for the jth class from M teacher models. Let $\mathcal{M}$ be the mechanism that produces a noisy max, $\arg\max_j\{n_j + Lap(2/\gamma)\}$. Then the privacy budget ϵ for a causal model is lower than that for the 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, $|n_j - n'_j| \leq 1$.

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,

$$n_i - 1 + r_i > n_j + 1 + r_j \Rightarrow n_i + r_i > n_j + 2 + r_j \tag{47}$$

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',

$$P(i|D') = P(r_i \geq r^* + 2) \qquad = 1 - 0.5\exp(\gamma)\exp(\frac{1}{2}\gamma r^*) \tag{48}$$

$$= 1 - \exp(\gamma)(1 - P(r_i \geq r^*)) \qquad = 1 - \exp(\gamma)(1 - P(i|D)) \tag{49}$$

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

$$\frac{P(i|D')}{P(i|D)} = \exp(\gamma) + \frac{1 - \exp(\gamma)}{P(i|D)} = \exp(\gamma) + \frac{1 - \exp(\gamma)}{P(r_i \geq r^*)} \leq \exp(\epsilon) \tag{50}$$

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

Below, we show that $|r^*|$ is higher for a causal model, and thus $P(r_i \geq r^*)$ is higher. Intuitively, $|r^*|$ is higher when there is more consensus between the M teacher models since $|r^*|$ is the difference between the votes for the highest voted class with the votes for the second-highest class.

Let us consider two causal teacher models $h1_c$ and $h2_c$, and two associational teacher models, $h1$ and $h2$. From Lemma 3, for any new x, and for same accuracies of the models, there is more consensus among causal models.

$$P(h1_c(x) = h2_c(x)) \geq P(h1(x) = h2(x)) \tag{51}$$

Hence $r^*_c \leq r^*$. From Equation 50, $\epsilon_c \leq \epsilon$.

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

$$P(i|D') = P(r_i \geq r^* + 2) \qquad = 0.5\exp(-\gamma)\exp(-\frac{1}{2}\gamma r^*) \tag{52}$$

$$= \exp(-\gamma)(P(r_i \geq r^*)) \qquad = \exp(-\gamma)(P(i|D)) \tag{53}$$

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

$$\frac{P(i|D')}{P(i|D)} = \exp(-\gamma) \tag{54}$$

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, $\epsilon_c \leq \epsilon$. □

C INFINITE SAMPLE ROBUSTNESS TO MEMBERSHIP INFERENCE ATTACKS

Corollary 2. *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.

$$\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} = \arg\min_{\mathtt{h} \in \mathcal{H}_{\mathtt{c}}} \mathcal{L}_{\mathtt{S} \sim \mathtt{P}}(\mathtt{h}, \mathtt{f}) = \arg\min_{\mathtt{h} \in \mathcal{H}_{\mathtt{c}}} \frac{1}{\mathtt{N}} \sum_{\mathtt{i}=1}^{\mathtt{N}} \mathtt{L}_{\mathtt{x_i}}(\mathtt{h}, \mathtt{f}) \tag{55}$$

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,

$$\lim_{\mathtt{N} \to \infty} \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} = \lim_{\mathtt{N} \to \infty} \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}'} \tag{56}$$

Equation 56 implies that as $N \to \infty$, 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 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 x is in the training set.

$$\begin{aligned}
\lim_{\mathtt{N} \to \infty} \mathtt{Adv}(\mathcal{A}, \mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) &= \Pr(\mathcal{A} = 1 | \mathtt{b} = 1) - \Pr(\mathcal{A} = 1 | \mathtt{b} = 0) \\
&= \mathbb{E}[\mathcal{A} | \mathtt{b} = 1] - \mathbb{E}[\mathcal{A} | \mathtt{b} = 0] = \mathtt{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) | \mathtt{b} = 1] - \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}}) | \mathtt{b} = 0] \\
&= \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}})] - \mathbb{E}[\mathcal{A}(\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}})] = 0
\end{aligned} \tag{57}$$

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

D ROBUSTNESS TO ATTRIBUTE INFERENCE ATTACKS

In addition to revealing membership in the training set, a model may also reveal the value of individual sensitive features of a test input, given partial knowledge of its features. For instance, given a training dataset of HIV patients, an adversary may infer other attributes of a person (e.g., genetic information) given that they know their demographics and other public features. As another example, it can be possible to infer a person's face based on hill climb on the output score for 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, e.g., using some demographics to predict disease, can be alleviated by causal models, since a feature will not be included in a model unless it directly affects the outcome.

Definition 7 (From Yeom et al. (2018))**.** *Let h be a model trained on a dataset $\mathcal{D}$(X,Y). Let A be an adversary with access to h, and a partial test input $x_A \subset x$. The attribute advantage of the adversary is the difference between true and false positive rates in guessing the value of a sensitive feature* $\mathtt{x_s} \notin \mathtt{x_A}$*. For a binary* $\mathtt{x_s}$*,*

$$\mathtt{Adv}(\mathcal{A}, \mathtt{h}) = \mathtt{Pr}(\mathcal{A} = 1 | \mathtt{x_s} = 1) - \mathtt{Pr}(\mathcal{A} = 1 | \mathtt{x_s} = 0) \tag{58}$$

Theorem 5. *Given a dataset $\mathcal{D}(X, Y)$ of size N and a structural causal network 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 $\mathtt{X_c}$.

$$\begin{aligned}
\mathtt{Adv}(\mathcal{A}, \mathtt{h}) &= \mathtt{Pr}(\mathcal{A} = 1 | \mathtt{x_s} = 1) - \mathtt{Pr}(\mathcal{A} = 1 | \mathtt{x_s} = 0) \\
&= \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1 | \mathtt{x_s} = 1) - \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1 | \mathtt{x_s} = 0) = \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1) - \mathtt{Pr}(\mathcal{A}(\mathtt{h}) = 1) = 0
\end{aligned} \tag{59}$$

□

E EXPERIMENTS

E.1 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 4. Note that the attacker model is trained on the confidence output of the target models.

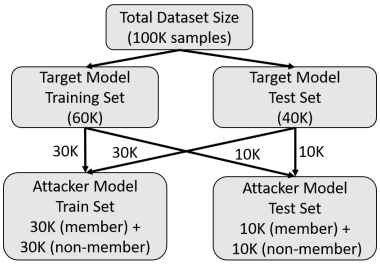


Figure 4: Dataset division for training target and attacker models.

E.2 FINE-GRAINED ATTACK ANALYSIS OF LEARNED CAUSAL MODELS.

We expand the evaluation of the attack accuracy for a learnt causal model on the `Sachs` dataset and `Akt` outcome. We increment the non-causal features to the learned model in addition to the true causal features. Table 2 shows the attack and prediction accuracy for this model when trained with the true causal, learned causal model with 1 and 2 non-causal features, and the results for the corresponding DNN model.

	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.*