Alleviating Privacy Attacks via Causal Learning

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

Abstract

Machine learning models, especially deep neural networks are known to be susceptible to privacy attacks such as *membership inference* where an adversary can detect whether a data point was used to train a model. Such privacy risks are exacerbated when a model is used for predictions 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 domains such as healthcare (Esteva et al., 2019) 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, 2019; Hayes et al., 2019). 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 used in sensitive applications such as in healthcare. For example, HIV-AIDS 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., 2018; Salem et al., 2019). 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. Generalizing to a new context, however, 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 such as neural networks 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 show that membership attacks are ineffective (almost a ran-

[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, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

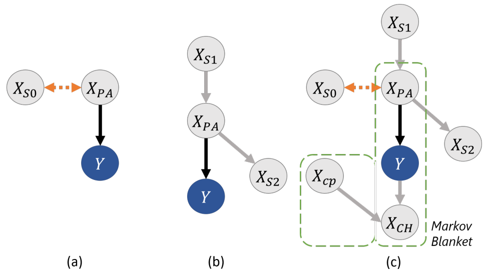


Figure 1: Structural causal model where a directed edge denotes a causal relationship; dashed bidirectional edges denote correlation. A causal predictive model includes only the parents of $\mathtt{Y}$: $\mathtt{X}_{\mathtt{PA}}$ [(a) and (b)]. Panel (c) shows the Markov Blanket of $\mathtt{Y}$.

dom 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 test distributions 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 predictive models generalize well since their output depends on stable 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 use score-based learning algorithms (Scutari, 2009), 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 different possible 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 a 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 (Pearl, 2009) states that a node is independent of its ancestors conditioned on all its parents. Hence $\mathtt{Y}$ 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 (Fig. 2). 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 domain (Fig. 2b), it will break the correlation between $\mathtt{Y}$ and

[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.

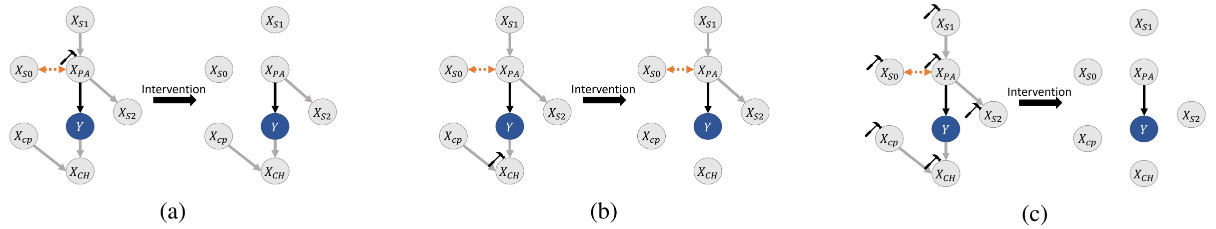


Figure 2: Interventions on (a) parents of $\mathtt{Y}$, (b) children of $\mathtt{Y}$, and (c) all features. The black hammer denotes an intervention and each right subfigure shows the resultant causal model. Relationship between causal features and $\mathtt{Y}$, $\mathtt{Y} = \mathtt{f}(\mathtt{X}_{\mathtt{PA}})$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

$\mathtt{X}_{\mathtt{CH}}$ and lead to incorrect predictions. To summarize, Fig. 2c demonstrates how a causal model based on parents is robust to all interventions on $\mathtt{X}$, 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, $\mathtt{L}(.,.)$ 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 to $\mathtt{h} : \mathtt{X} \to \mathtt{Y}$ as the hypothesis function or simply the model. Then, $\mathtt{L}(\mathtt{h},\mathtt{h}')$ is a loss function quantifying the difference between any two models $\mathtt{h}$ and $\mathtt{h}'$.

Definition 1. In-Distribution Generalization Error (IDE). *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 in-distribution generalization error is given by,*

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

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,*

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

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

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

Intuitively, the term $\mathtt{disc}_{\mathtt{L},\mathcal{H}}(\mathtt{P},\mathtt{P}^*)$ denotes the distance between two distributions. Higher the distance, higher is the chance of an error when transferring model $\mathtt{h}$ from one distribution to another. Using these definitions, next we characterize the generalization property of causal models.

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

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

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* $\mathtt{ODE\text{-}Bound}$*) for a causal model* $\mathtt{h}_{\mathtt{c}} \in \mathcal{H}_{\mathtt{C}}$ *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* $(1-\delta)^2$*.*

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

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

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

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

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

Proof Sketch. As an example, consider a colored MNIST data distribution $\mathtt{P}$ such that the true label $\mathtt{Y}$ is assigned based on the shape of a digit. Here an input's shape features represent the causal features ($\mathtt{X_C}$). If the shape is closest to shapes for $\{0,1,2,3,4\}$ then $\mathtt{Y}=\mathtt{0}$, else $\mathtt{Y}=\mathtt{1}$. 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 $\mathtt{P}(\mathtt{Y}|\mathtt{X_C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X_C})$ 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, $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P}} = \mathtt{h}^{\mathtt{OPT}}_{\mathtt{c,P^*}}$. However for associational models, the optimal models may not be the same $\mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P}} \neq \mathtt{h}^{\mathtt{OPT}}_{\mathtt{a,P^*}}$ 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 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* $\mathtt{h}^{\mathtt{min}}_{\mathtt{c,S}} : \mathtt{X_C} \to \mathtt{Y}$ *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 the causal model on such* $\mathtt{x}'$ *is less than or equal to that for the associational model.*

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

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

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

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

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

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

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

We now prove our main result on differential privacy.

Theorem 3. *Let* $\hat{\mathcal{F}}_c$ *and* $\hat{\mathcal{F}}_a$ *be two 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,* $\mathtt{Lap}(\mathtt{Z})$, *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 any 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* $\mathtt{n}_\mathtt{k}$ *be the votes for the* $\mathtt{k}$*th class from* $\mathtt{M}$ *teacher models. Let* $\mathcal{M}$ *be the mechanism that produces a noisy max,* $\arg\max_\mathtt{k}\{\mathtt{n}_\mathtt{k} + \mathtt{Lap}(2/\gamma)\}$. *Then the privacy budget* $\epsilon_\mathtt{c}$ *for a causal model trained on* $\mathtt{D}$ *is lower than that for an associational model.*

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., 2019). 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 an 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., 2019). 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., 2019).

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., 2018). 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 allows the adversary to detect differences in losses for the test versus training datasets. As we show, 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* $\mathtt{S}(\mathtt{X},\mathtt{Y}) \sim \mathtt{P}$ *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* $\mathtt{P}(\mathtt{Y}|\mathtt{X}_\mathtt{C}) = \mathtt{P}^*(\mathtt{Y}|\mathtt{X}_\mathtt{C})$. *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., 2019).

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* $\mathtt{1} - \mathtt{L}_\mathtt{x}(\mathtt{h},\mathtt{y})/\mathtt{B}$.

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 a data point belongs to the train dataset $\mathtt{S}$. The membership advantage of the bounded loss adversary $\mathcal{A}_\mathtt{BL}$ for any model $\mathtt{h}$ trained on dataset $\mathtt{S} \sim \mathtt{P}$ is given by,

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

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,

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

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:

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

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

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.

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

□

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, it follows 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 Sketch. 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. Detailed proof is in Suppl. Section D.

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 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* $\mathtt{x_s}$*,*

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

Theorem 7. *Given a dataset* $\mathtt{S(X, Y)}$ *of size* $\mathtt{n}$ *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 evaluate on two types of datasets: 1) Four datasets generated from known Bayesian Networks and 2) Colored images of digits from the MNIST dataset. Code is available at `https://github.com/microsoft/robustdg`.

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 $178 - 10$k 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 $\mathtt{P(X_i|Parents(X_i))}$; 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 dataset where it is difficult to construct a causal graph of the input features. For this, we consider 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 from Arjovsky et al. (2019) to generate the dataset and perform our evaluation[4].

4.1 Results for Bayesian Network Datasets

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. The causal and DNN model are the *targets* on which the attack is perpetrated. 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., 2019). 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 Fig. 5 in the Suppl. Section G. Note that the reported attack accuracies are an upper bound since we assume that the adversary has access to a subset of training data for the ML model.

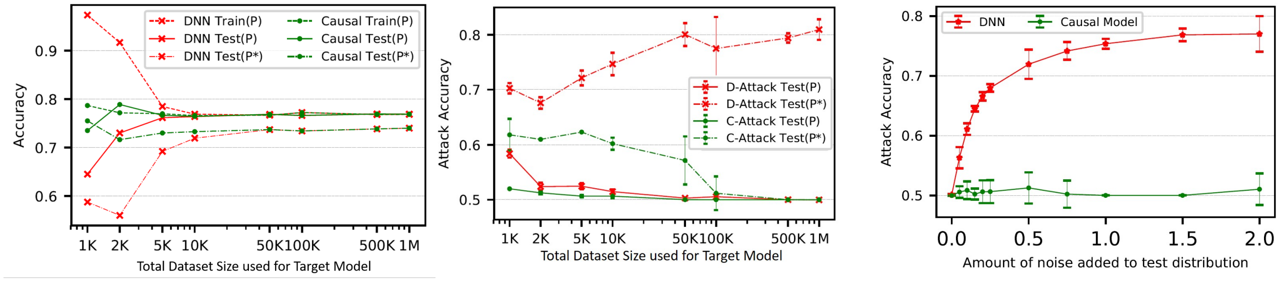
To train the causal model, we use the bnlearn library in R language that supports maximum likelihood estimation of the parameters in Y's conditional probability table. For prediction, we use the `parents` method to predict the class of any specific variable. To train the DNN model and the attacker model, 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 on sample sizes ranging from 1K to 1M data points. We use two test datasets: Test(P) is drawn from the same distribution as the training data and Test(P*) is drawn 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 $\mathtt{Pr(X)}$ uniformly at random (later, we add noise to the original value). Our goal with generating Test (P*) is to capture extreme shifts in the distribution for input features.

[2] www.bnlearn.com/bnrepository

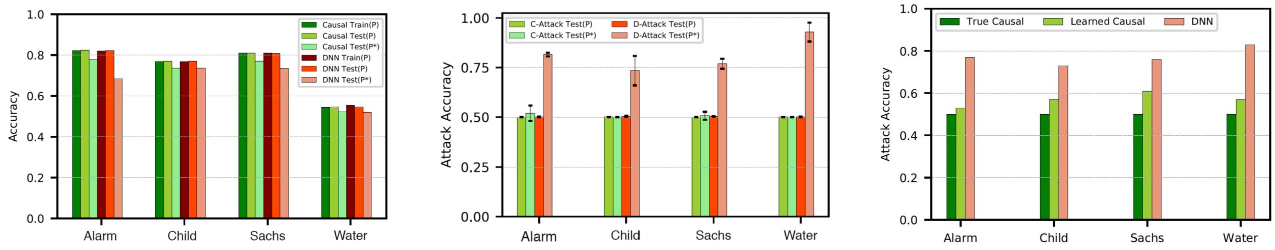
[3] http://yann.lecun.com/exdb/mnist/

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



(a) (b) (c)

Figure 3: Results for the 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: Models trained on a dataset of size of 60K for all four Bayesian Networks. (a) is the accuracy of the target model, (b) is the attack accuracy against 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. Fig. 3a shows the target model accuracies 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 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. Fig. 4a shows similar results for the accuracy on all four 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. Fig. 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 (Fig. 3a). These results show that causal models generalize better than DNNs across input distributions.

Fig. 4b shows a similar result for all four datasets. While the attack accuracy for DNNs and the causal model is close to 0.5 for Test(P) dataset, the attack accuracy is significantly higher for DNNs than causal model for Test(P*) dataset. This empirically confirms our claim that 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 $\Pr(\mathtt{X})$ changes, we generate test data from different distributions by adding varying amount of noise to the true probabilities. We range

	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.

the noise value between 0 to 2 and add it to the individual probabilities which are then normalized to sum up to 1. Fig. 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 its 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 outputs 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 Fig. 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 increasing 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

Arjovsky et al. (2019) proposed a way to train a causal model by minimizing the loss across different environments or data distributions. Using this approach, we train an invariant risk minimizer (IRM) and an empirical risk minimizer (ERM) model on the colored MNIST data. Since IRM learns a feature representation such that the same model is optimal for the two training domains, it aims to learn the causal features (shape) that are 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. The causal model has attack accuracy close to a random guess whereas 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) demonstrated 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., 2019), text prediction generative models (Carlini et al., 2019; Song & Shmatikov, 2019) and federated learning models (Nasr et al., 2018). 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., 2018) 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). Therefore, 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. (2016) proposed a method to privately reveal parameters of a causal model using the framework of differential privacy. Instead of a specific causal algorithm, our focus is on the privacy benefits of causal learning for general predictive tasks, complementing recent work on using causality to enhance properties such as explainability (Datta et al., 2016) or fairness (Kusner et al., 2017) of machine learning models.

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., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

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. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

Kusner, M. J., Sun, Y., Sridharan, K., and Weinberger, K. Q. Private causal inference. In *Artificial Intelligence and Statistics*, pp. 1308–1317, 2016.

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.

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. 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, 2018.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning. In *2019 IEEE symposium on security and privacy*, 2019.

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. In *NDSS*, 2019.

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. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 196–206, 2019.

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.