Private Causal Inference

Matt J. Kusner
Washington University in St. Louis
mkusner@wustl.edu

Yu Sun
Cornell University
ys646@cornell.edu

Karthik Sridharan
Cornell University
sridharan@cs.cornell.edu

Kilian Q. Weinberger
Cornell University
kqw4@cornell.edu

Abstract

Causal inference deals with identifying which random variables "cause" or control other random variables. Recent advances on the topic of causal inference based on tools from statistical estimation and machine learning have resulted in practical algorithms for causal inference. Causal inference has the potential to have significant impact on medical research, prevention and control of diseases, and identifying factors that impact economic changes to name just a few. However, these promising applications for causal inference are often ones that involve sensitive or personal data of users that need to be kept private (e.g., medical records, personal finances, etc). Therefore, there is a need for the development of causal inference methods that preserve data privacy. We study the problem of inferring causality using the current, popular causal inference framework, the additive noise model (ANM) while simultaneously ensuring privacy of the users. Our framework provides differential privacy guarantees for a variety of ANM variants. We run extensive experiments, and demonstrate that our techniques are practical and easy to implement.

1 Introduction

Causal identification allows one to reason about how manipulations of certain random variables (the causes) affect the outcomes of others (the effects). Uncovering these causal structures has implications ranging from creating government policies to informing health-care practices. Causal inference was motivated by the impossibility of randomized intervention experiments in many cases, and the ambiguity of conditional independence testing [32, 25]. In the absence of interventions, it attempts to discover the underlying causal relationships of a set of random variables entirely based on samples from their joint distribution. The field of causal inference is now a mature research area, covering learning topics as diverse as supervised batch inference [19, 23, 26], time-series causal prediction [10], and linear dynamical systems [30]. Many inference methods require only a regression technique and a way to compute the independence between two distributions given samples [13, 16].

One would hope that researchers could publicly release their causal inference findings to inform individuals and policy makers. One of the primary roadblocks to doing so is that often causal inference is performed on data that individuals may wish to keep private, such as data in the fields of medical diagnosis, fraud detection, and risk analysis. Currently, no causal inference method has formal guarantees about the privacy of individual data, which may be able to be inferred via attacks such as reconstruction attacks [3].

Arguably one of the best notion of privacy is differential privacy, introduced by Dwork et al. [6] and since used throughout machine learning [4, 15, 20, 2, 5]. Differential privacy guarantees that the outcome of an algorithm only reveals aggregate information about the entire dataset and never about the individual. An individual who is considering to participate in a study can be reassured that his/her personal information cannot be recovered with extremely high probability.

To our knowledge, this paper is the first to investigate private causal inference. We show that it is possible to privately release the quantities produced by the highly-successful additive noise model (ANM) framework by adding small amounts of noise, as dictated by differential privacy. Furthermore, these private quantities, with high probability, do not change the causal inference result, so long as it is confident enough. We demonstrate on a set of real-world causal inference datasets how our privacy-preserving methods can be readily and usefully applied.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

2 Related Work

Discovering the causal nature between random events has captivated researchers and philosophers long before the formal developments of statistics. This interest was formalized by Reichenbach & Reichenbach [28] who argued that *all* statistical correlations in data arise from underlying causal structures between the concerned random variables. For example, the correlation between smoking and lung cancer was found to arise from a direct causal link [7].

One of the most popular causal inference alternatives to conditional independence testing is the Additive Noise Model (ANM) approach developed by Hoyer et al. [13] and used in many recent works [35, 21, 18, 1]. ANMs, originally designed for inferring whether $X \rightarrow Y$ or $Y \rightarrow X$ and later extended to large numbers of random variables, work under the assumption that the effect is a non-linear function of the cause plus independent noise. ANMs are one of many proposed causal inference methods in recent literature [16, 9, 19, 29]

Work by Spirtes et al. [32], Pearl [25] shows how to determine if $X \rightarrow Y$ when these variables are a part of a larger 'causal network', via conditional independence testing. One downside to conditional independence based approaches is that inherently they cannot distinguish between Markov-equivalent graphs. Thus it may be possible that a certain set of conditional independences imply both $X \rightarrow Y$ and $Y \rightarrow X$. Furthermore, if X and Y are the only variables in the causal network there is no conditional independence test to determine whether $X \rightarrow Y$ or $Y \rightarrow X$.

3 Background

Our aim is to protect the privacy of individuals who submit personal information about two random variables of interest X and Y. Their information should remain private when it is used to infer whether X causes Y ($X \rightarrow Y$), or Y causes X ($Y \rightarrow X$) using the ANM framework. This personal information comes in the form of i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from the joint distribution $\mathbb{P}_{X,Y}$. We will assume that, 1. There is no confounding variable Z that commonly causes or is a common effect of X and Y. 2. X and Y do not simultaneously cause each other.

3.1 Additive Noise Model

Deciding on the causal direction between two variables X and Y from a finite sample set has motivated an array of research [8, 17, 33, 13, 35, 22, 16, 18, 19]. Perhaps one of the most popular results is the Additive Noise Model (ANM) proposed by Hoyer et al. [13]. The ANM framework assumption is defined as follows.

Definition 1. *Two random variables X, Y with joint density $p(x, y)$ are said to 'satisfy an ANM' $X \rightarrow Y$ if there exists a non-linear function $f : \mathcal{R} \rightarrow \mathcal{R}$ and a random noise variable N_Y, independent from X, i.e. $X \perp\!\!\!\perp N_Y$, such that*

$$Y = f(X) + N_Y.$$

As defined, an ANM $X \rightarrow Y$ implies a functional relationship mapping X to Y, alongside independent noise. In order for this model to be useful for causal inference we would like the induced joint distribution $\mathbb{P}_{X,Y}$ for this ANM to be somehow identifiably different from the one induced by the ANM $Y \rightarrow X$. If so, we say that the causal direction is *identifiable* [23]. If not, we have no hope of recovering the causal direction purely from samples under the ANM.

Hoyer et al. [13] showed that ANMs are generically identifiable from i.i.d. samples from $\mathbb{P}_{X,Y}$ (except for a few special cases of non-linear functions f and noise distributions). The intuition behind this is for the $X \rightarrow Y$ ANM, consider for most non-linear f and (for simplicity) 0-mean N_Y, the density $p(y|x)$ has mean $f(x)$ with distribution given by N_Y. This implies that $p(y - f(x)|x)$ has distribution N_Y that is independent of X. However, $p(x - f^{-1}(y)|y)$ is for many choices of f and N_Y not independent of y.

Algorithm 1 ANM Causal Inference [23]

Input: train/test data $\{x_i, y_i\}_{i=1}^n$, $\{x'_i, y'_i\}_{i=1}^m$
Regress on training data, to yield $\hat{f}, \hat{g}$, such that:
$\hat{f}(x_i) \approx y_i, \quad \hat{g}(y_i) \approx x_i, \ \forall i$
Compute residuals on test data:
$\mathbf{r}'_Y := \mathbf{y}' - \hat{f}(\mathbf{x}'), \quad \mathbf{r}'_X := \mathbf{x}' - \hat{g}(\mathbf{y}')$
Calculate dependence scores:
$s_{X \rightarrow Y} := s(\mathbf{x}', \mathbf{r}'_Y), \quad s_{Y \rightarrow X} := s(\mathbf{y}', \mathbf{r}'_X)$
Return: $s_{X \rightarrow Y}$, $s_{X \rightarrow Y}$, and D, where
$\mathrm{D} = \begin{cases} X \rightarrow Y & \text{if } s_{X \rightarrow Y} < s_{Y \rightarrow X} \\ Y \rightarrow X & \text{if } s_{X \rightarrow Y} > s_{Y \rightarrow X} \end{cases}$

3.2 Inferring Causality

Mooij et al. [23] give a practical algorithm for determining the causal relationship between X and Y (i.e., either $X \rightarrow Y$ or $Y \rightarrow X$), as shown in Algorithm 1. The first step is to partition the i.i.d. samples into a training and a testing set. We use the training set to train the regression functions $\hat{f} : X \rightarrow Y$ and $\hat{g} : Y \rightarrow X$. We use the testing set to compute the residuals $\mathbf{r}'_Y = \mathbf{y}' - \hat{f}(\mathbf{x}')$ and $\mathbf{r}'_X := \mathbf{x}' - \hat{g}(\mathbf{y}')$. If we have an ANM $X \rightarrow Y$ then the residual $\mathbf{r}'_Y$ is an estimate of the noise N_Y which is assumed to be independent of X. Therefore, we calculate the dependence between the residual $\mathbf{r}'_Y$ and the input $\mathbf{x}'$, $s_{X\to Y} := s(\mathbf{x}', \mathbf{r}'_Y)$, and $s_{Y\to X} := s(\mathbf{y}', \mathbf{r}'_X)$, using a dependence score $s(\cdot,\cdot)$. If $s_{X\to Y}$ is less than $s_{Y\to X}$, then we declare $X\to Y$, otherwise $Y\to X$.

3.3 Dependence Scores

Crucially, the ANM approach hinges on the choice of dependence score $s(\cdot,\cdot)$. There have been many proposals, and we give a quick review of the most popular methods (for a detailed review see Mooij et al. [23]).

Spearman's ρ is a rank correlation coefficient that describes the extent to which one random variable is a monotonic function of the other. Specifically, imagine independently sorting the observations $\{a_1,\ldots,a_m\}$ and $\{b_1,\ldots,b_m\}$ by value in increasing order. Let o_i^a be the rank of a_i in the a-ordering, and similarly, o_i^b for b_i in the b-ordering. Then Spearman's ρ is,

$$s(\mathbf{a},\mathbf{b}) := \left|1 - \frac{6\sum_{i=1}^m d_i^2}{m(m^2-1)}\right|$$

where $d_i := (o_i^a - o_i^b)$ are the rank differences for $\mathbf{a},\mathbf{b}$.

Kendall's τ**.** Similar to Spearman's ρ, the Kendall τ rank score calls a pair of indices (i,j) *concordant* if it is the case that $a_i > a_j$ and $b_i > b_j$. Otherwise (i,j) is called *discordant*. Then the dependence score is defined as

$$s(\mathbf{a},\mathbf{b}) := \frac{|C-D|}{\frac{1}{2}m(m-1)}$$

where C is the number of concordant pairs and D is the number of discordant pairs.

HSIC Score. The first proposed score for the ANM causal inference is based on the Hilbert-Schmidt Independence Criterion (HSIC) [11], which was used by Hoyer et al. [13]. They compute an estimate of the p-value of the HSIC under the null hypothesis of independence, selecting the causal direction having the lower p-value. Alternatively, one can use an estimator to the HSIC value itself:

$$s(\mathbf{a},\mathbf{b}) := \widehat{\mathrm{HSIC}}_{k_{\theta(\mathbf{a})},k_{\theta(\mathbf{b})}}(\mathbf{a},\mathbf{b}) \qquad (1)$$

where k_θ is a kernel with parameters θ. Mooij et al. [23] show that under certain assumptions the algorithm in section 1 with the HSIC dependence score is consistent for estimating the causal direction in an ANM.

Variance Score. When the noise variables in the ANM are Gaussian, the variance score was proposed in Bühlmann et al. [1], and defined as $s(\mathbf{a},\mathbf{b}) := \log\mathbb{V}(\mathbf{a}) + \log\mathbb{V}(\mathbf{b})$. Changes to a single input value can induce arbitrarily large changes to this score, which makes the variance score ill suited to preserve differential privacy.

IQR Score. We introduce a robust version of this score by replacing the variance of the random variables with their interquartile range (IQR). The IQR is the difference between the third and first quartiles of the distribution and can be estimated empirically. We defined the following IQR-based score:

$$s(\mathbf{a},\mathbf{b}) := \log\mathrm{IQR}(\mathbf{a}) + \log\mathrm{IQR}(\mathbf{b}). \qquad (2)$$

3.4 Differential Privacy

We assume that the data set $\mathcal{D} = \{(x_i,y_i)\}$ contains sensitive data that should not be inferred from the release of the dependence scores. One of the most widely accepted mechanisms for private data release is *differential privacy* [6]. In a nutshell it ensures that the released scores can only be used to infer aggregate information about the data set and never about an individual datum (x_i,y_i).

Let us define the Hamming distance between two data sets $d_H(\mathcal{D},\tilde{\mathcal{D}})$ between two data sets $\mathcal{D}$ and $\tilde{\mathcal{D}}$ as the number of elements in which these two sets differ. If a data set $\mathcal{D}$ is changed to $\tilde{\mathcal{D}}$, a distance $d_H(\mathcal{D},\tilde{\mathcal{D}}) \le 1$ implies that at most one element was added, removed, or substituted.

Definition 2. *A randomized algorithm $\mathcal{A}$ is (ϵ,δ)-* ***differentially private*** *for $\epsilon,\delta \ge 0$ if for all $\mathcal{O} \in$* Range$(\mathcal{A})$ *and for all neighboring datasets $\mathcal{D},\tilde{\mathcal{D}}$ with $d_H(\mathcal{D},\tilde{\mathcal{D}}) \le 1$ we have that*

$$\Pr\big[\mathcal{A}(\mathcal{D}) = \mathcal{O}\big] \le e^\epsilon \Pr\big[\mathcal{A}(\tilde{\mathcal{D}}) = \mathcal{O}\big] + \delta. \qquad (3)$$

One of the most popular methods for making an algorithm $(\epsilon,0)$-differentially private is the Laplace mechanism [6]. For this mechanism we must define an intermediate quantity called the **global sensitivity**, $\Delta_\mathcal{A}$ describing how much $\mathcal{A}$ changes when $\mathcal{D}$ changes,

$$\Delta_\mathcal{A} := \max_{\mathcal{D},\tilde{\mathcal{D}}\subseteq\mathcal{X} \text{ s.t. } d_H(\mathcal{D},\tilde{\mathcal{D}})\le 1} |\mathcal{A}(\mathcal{D}) - \mathcal{A}(\tilde{\mathcal{D}})|.$$

The Laplace mechanism hides the output of $\mathcal{A}$ with a small amount of additive random noise, large enough to hide the impact of any *single* datum (x_i,y_i).

Definition 3. *Given a dataset $\mathcal{D}$ and an algorithm $\mathcal{A}$, the* ***Laplace mechanism*** *returns $\mathcal{A}(\mathcal{D})+\omega$, where ω is a noise variable drawn from* $\mathrm{Lap}(0,\Delta_\mathcal{A}/\epsilon)$*, the Laplace distribution with scale parameter $\Delta_\mathcal{A}/\epsilon$.*

It may be that the global sensitivity of an algorithm $\mathcal{A}$ is unbounded in general, but can be bounded in the context of a specific data set $\mathcal{D}$ over all neighbors $\tilde{\mathcal{D}}$. For such datasets we can bound the **local sensitivity**

$$\Delta(\mathcal{D})_\mathcal{A} := \max_{\tilde{\mathcal{D}}\subseteq\mathcal{X} \text{ s.t. } d_H(\mathcal{D},\tilde{\mathcal{D}})\le 1} |\mathcal{A}(\mathcal{D}) - \mathcal{A}(\tilde{\mathcal{D}})|.$$

Table 1: Dependence scores and their privacy. A checkmark indicates that there exist meaningful bounds on either the global or local sensitivity.

Score	Test Global Sense.	Test Local Sense.	Training Global Sense.	Training Local Sense.
Spearman's ρ	✓	✓	-	✓
Kendall's τ	✓	✓	-	✓
HSIC	✓	✓	✓	✓
IQR	-	✓	-	✓

If an algorithm has bounded global sensitivity it certainly has bounded local sensitivity. Nissim et al. [24], Dwork & Lei [4], Jain & Thakurta [14] show how to use the local sensitivity to cleverly produce private quantities for datasets with bounded local sensitivity.

4 Test Set Privacy

The data is partitioned into training and test set, which are used in different ways. We therefore introduce mechanisms to preserve training and test set privacy respectively, which can be used jointly. Specifically, we show how to privatize the dependence scores $s_{X\to Y}, s_{Y\to X}$. The reason for this is four-fold: 1. Privatizing the dependence score immediately privatizes the causal direction D, because operations on differentially private outputs preserve privacy (so long as they do not touch the data). 2. Releasing the scores indicates how confident the ANM method is about the causal direction, which is absent from the binary output D. 3. It is unclear which dependence score is best for a particular dataset, so we privatize multiple scores and leave this choice to the practitioner. In this section we begin with test set privacy and describe training set privacy in Section 5. Table 1 gives an overview of test and training set privacy results for the dependence scores that we consider.

Let $(\mathbf{x}', \mathbf{y}')$ be the initial test data and $(\tilde{\mathbf{x}}', \tilde{\mathbf{y}}')$ be the test data after a single change in the dataset. Let $\tilde{\mathbf{x}}' = [x'_1, \ldots, x'_{k-1}, \tilde{x}'_k, x'_{k+1}, \ldots, x'_m]^\top$ and similarly for $\tilde{\mathbf{y}}$ so that this single change occurs at some index k. The key to preserving privacy is to show that the selected dependence score $s(\cdot,\cdot)$ can be privatized. We show that if our dependence score is a rank correlation coefficient (Spearman's ρ, Kendall's τ) or the HSIC score [11], we can readily bound its test set global sensitivity when applied to $(\mathbf{x}', \mathbf{y}')$ versus $(\tilde{\mathbf{x}}', \tilde{\mathbf{y}}')$. As the IQR score has bounded test set local sensitivity we can apply the algorithm of Dwork & Lei [4] for privacy.

4.1 Rank Correlation Coefficients

We first demonstrate global sensitivity for the two rank correlation scores in Section 3.

Theorem 1. *The rank correlation coefficients have the following global sensitivities,*

1. *Let $\rho(\cdot,\cdot)$ be Spearman's ρ score, then*

$$|\rho(\mathbf{x}', \mathbf{r}'_Y) - \rho(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)| \leq \frac{30}{m}$$

2. *Let $\tau(\cdot,\cdot)$ be Kendall's τ score, then*

$$|\tau(\mathbf{x}', \mathbf{r}'_Y) - \tau(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)| \leq \frac{4}{m}$$

Proof. Our goal is to bound the following global sensitivity in both scores: $|s(\mathbf{x}', \mathbf{r}'_Y) - s(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)|$. For Spearman's ρ, suppose the change is on a_k and b_k, it is easy to verify that 1) d_i changes by at most 2, for $i \neq k$; 2) d_k changes by at most $m-1$; 3) $d_i \leq m-1$ for all i. Since $d_i^2 - (d_i - 2)^2 = 4(d_i - 1) \leq 4(m-2)$ for $i \neq k$, the maximum change inside the summation is upper bounded by $(m-1)(4m-8) + (m-1)^2$. Therefore, global sensitivity of ρ is bounded by

$$\frac{6(m-1)(5m-3)}{m(m^2-1)} \leq \frac{30}{m}$$

.

For Kendall's τ we can affect at most $(m-1)$ pairs by moving a single element of $\mathbf{x}'$, as well as $(m-1)$ pairs for changing $\mathbf{r}'_Y$ (either from concordant pairs to discordant pairs, or vice versa). Therefore, the global sensitivity of Kendall's τ is

$$|s(\mathbf{x}', \mathbf{r}'_Y) - s(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)| \leq \frac{2(m-1)}{\frac{1}{2}m(m-1)} \leq \frac{4}{m}$$

■

The bound on the global sensitivity Δ of our scores enables us to apply the Laplace mechanism [6] to produce $(2\epsilon, 0)$-differentially private scores: $p_{X\to Y}, p_{Y\to X}$. Specifically, we add Laplace noise $\text{Lap}(0, \Delta/\epsilon)$ to our Spearman's ρ and Kendall's τ scores to preserve privacy w.r.t. the test set. Moreover, as a general property of differential privacy we can compute any functions on these private scores and, so long as they do not touch the data, the outputs of these functions are also private. This means that we can compute the inequality $p_{X\to Y} < p_{Y\to X}$ to decide if X causes Y or vice-versa privately.

An important consideration is to what degree the addition of noise affects the true decision: $s_{X\to Y} < s_{Y\to X}$. Importantly, we can prove that, in certain cases, the addition of Laplace noise required by the mechanism is small enough to not change the direction of causal inference. These are cases in which there is a large 'margin' between the scores $s_{X\to Y}$ and $s_{Y\to X}$. So long as this margin is large enough and in the correct order the addition of Laplace noise has no effect on the inference with high probability.

Theorem 2. *Given two random variables X, Y who have w.l.o.g. the causal relationship $X \rightarrow Y$, assume that they produce correctly-ordered scores: $s_{X\rightarrow Y} < s_{Y\rightarrow X}$, with margin $\gamma = s_{Y\rightarrow X} - s_{X\rightarrow Y}$. Let $p_{X\rightarrow Y}, p_{Y\rightarrow X}$ be these scores after applying the Laplace mechanism [6] with scale $\sigma = \Delta/\epsilon$ then the probability of correct inference with these private scores is,*

$$\mathbb{P}(p_{X\rightarrow Y} < p_{Y\rightarrow X}) = 1 - \frac{\gamma + 2\sigma}{4\sigma} e^{-\frac{\gamma}{\sigma}}.$$

We leave the proof to the appendix. Note that the probability of incorrect inference decreases nearly exponentially as the margin γ increases. This is a particularly nice property as the margin essentially describes the confidence of the (non-private) causal inference prediction: large γ corresponds to high confidence in the inference. Additionally, there is an exponential decrease as m and ϵ grow. In section 6, we show on real-world causal inference data that we can accurately recover the true causal direction for a variety ϵ settings.

4.2 HSIC Score

We begin by defining the empirical estimate of the HSIC score given kernels k, l:

$$\widehat{\text{HSIC}}_{k,l}(\mathbf{x}', \mathbf{r}'_Y) := \frac{1}{(m-1)^2}\text{trace}(KHLH) \tag{4}$$

where $K_{ij} = k(x'_i, x'_j)$, $L_{ij} = l(r'_{Y,i}, r_{Y,j})$ and $H_{ij} = \delta_{\{i=j\}} - 1/m$. We assume k, l are bounded above by 1 (e.g., the squared exponential kernel, the Matern kernel [27]). Our goal is to show that when we replace $(\mathbf{x}', \mathbf{y}')$ with $(\tilde{\mathbf{x}}', \tilde{\mathbf{y}}')$ the global sensitivity is small. Specifically we prove the following theorem.

Theorem 3. *The score in eq. (4) has a global sensitivity of at most $\frac{16m-8}{(m-1)^2}$. Specifically,*

$$|\widehat{HSIC}_{k,l}(\mathbf{x}', \mathbf{r}'_Y) - \widehat{HSIC}_{k,l}(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)| \leq \frac{16m-8}{(m-1)^2}$$

Proof. For simplicity define $\mathcal{H}(\cdot,\cdot) := \widehat{\text{HSIC}}_{k,l}(\cdot,\cdot)$. Note that, as the trace is cyclic: $\text{trace}(KHLH) = \text{trace}(HKHL)$. Further, let $\tilde{K}, \tilde{L}$ be the kernels defined on the modified data $(\tilde{\mathbf{x}}', \tilde{\mathbf{y}}')$. Then as the data is represented purely through the kernel matrices and the trace is Lipschitz w.r.t. these matrices, we can apply the triangle inequality to yield,

$$|\mathcal{H}(\mathbf{x}', \mathbf{r}'_Y) - \mathcal{H}(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)| \leq \frac{\|HLH\|_\infty \|K - \tilde{K}\|_1}{(m-1)^2} + \frac{\|HKH\|_\infty \|L - \tilde{L}\|_1}{(m-1)^2}$$

To bound the infinity norms, let $\overline{L} = HLH$, then

$$|\overline{L}_{ij}| = \left| L_{ij} - \frac{\sum_{a=1}^m L_{aj}}{m} - \frac{\sum_{b=1}^m L_{ib}}{m} + \frac{\sum_{a,b=1}^m L_{ab}}{m^2} \right| \leq 4$$

as $L_{ij} \leq 1$ (this inequality also holds for HKH). Finally, note that as there is only a single-element difference between $(\mathbf{x}', \mathbf{r}'_Y)$ and $(\tilde{\mathbf{x}}', \tilde{\mathbf{r}}'_Y)$, we have that $\|K - \tilde{K}\|_1 \leq 2m - 1$ (and also for $L, \tilde{L}$). ■

In fact, we can improve this bound to $\frac{12m-11}{(m-1)^2}$ using trace identities. We leave the proof of this to the appendix. Given this global sensitivity bound we can use Theorem 2 to guarantee that under certain conditions the Laplace mechanism w.h.p. does not change the direction of causal influence.

4.3 IQR Score

Unfortunately the IQR does not have a bounded global sensitivity, as there exist datasets for which the IQR can change by an unbounded amount. Instead, Dwork & Lei [4] offer an efficient technique to privately release the IQR. We give a slightly modified version of their Algorithm in the appendix.

First the algorithm defines two intervals B_1 and B_2 which both contain $\text{IQR}(\mathbf{X})$. If the IQR were to be pushed out of both of these intervals it would imply that the IQR changed by a factor of e. Therefore we loop over both intervals and calculate the number of points A_j that an adversary would need to change to push the IQR out of B_1 or B_2. Note that A_j is itself a data-sensitive query and so, to preserve privacy of this query, we can add Laplace noise to it. Then, if one of these noisy estimates $R_j = A_j + z$, where $z \sim \text{Lap}(0, 1/\epsilon)$ is larger than some threshold, it implies that with high probability (exactly $1-\delta$), that the $\text{IQR}(\mathbf{X})$ has multiplicative sensitivity of at most e, *for the specific dataset* $\mathbf{X}$. Note that this is precisely the local sensitivity as defined in Section 3, as it is specific to $\mathbf{X}$. This means that we can add Laplace noise z to $\log \text{IQR}(\mathbf{X})$. If neither of the R_j are above the threshold then the algorithm returns null: $\perp$. This algorithm was shown to be $(3\epsilon, \delta)$-differentially private.

In our case we would like to release four private IQR scores. Note that we must look at $\mathbf{x}'$ three separate times: for $\text{IQR}(\mathbf{x}'), \text{IQR}(\mathbf{r}'_Y)$, and $\text{IQR}(\mathbf{r}'_X)$ (and three times as well for $\mathbf{y}'$). Therefore for both $\mathbf{x}'$ and $\mathbf{y}'$ we are composing three differentially private outputs. Under simple composition this would lead to $(9\epsilon, 3\delta)$ differential privacy for both $\mathbf{x}'$ and $\mathbf{y}'$. However, we can make use of Corollary 3.21 in Dwork & Roth [5] to give $(\epsilon', 3\delta + \delta')$-differential privacy, for $0 < \epsilon' < 1$ and

$\delta' > 0$, over three repeated mechanisms by ensuring each private mechanism is $(3\epsilon, \delta)$-private, where $3\epsilon = \epsilon'/(2\sqrt{6\log(1/\delta')})$.

The remaining question is whether this noise addition causes one to infer the incorrect causal direction. Again, as long as there is a significant margin between the scores, we can preserve the correct causal inference with high probability as follows.

Theorem 4. *Let $Q_{\mathbf{x}'} = \log IQR(\mathbf{x}')$, and similarly for $Q_{\mathbf{y}'}, Q_{\mathbf{r}'_X}, Q_{\mathbf{r}'_Y}$, be the true log-IQR scores. As well let $P_{\mathbf{x}'}, P_{\mathbf{y}'}, P_{\mathbf{r}'_X}, P_{\mathbf{r}'_Y}$ be the private versions, multiplied by e^z noise where $z \sim Lap(0, 1/\epsilon)$. The the following results hold:*

1. *[4] If the number of data-points needed to significantly change the IQR, A_j, is less than e then, the probability that any one of the private IQR P_* is released is small:*

$$\mathbb{P}\left[P_* \neq \perp | A_1 \text{ or } A_2 \leq e\right] \leq \frac{3\delta}{2}.$$

2. *If all private log-IQR scores are released, and the relationship between the true scores holds $Q_{\mathbf{x}'} + Q_{\mathbf{r}'_Y} < Q_{\mathbf{y}'} + Q_{\mathbf{r}'_X}$ (which implies $X \rightarrow Y$), then the probability that we make the correct causal inference from the private scores is large,*

$$\mathbb{P}[P_{\mathbf{x}'} + P_{\mathbf{r}'_Y} < P_{\mathbf{y}'} + P_{\mathbf{r}'_X}] =$$
$$1 - \frac{e^{\frac{-\gamma}{\sigma}}}{96\sigma^3}\left(48\sigma^3 + 33\sigma^2\gamma + 9\sigma\gamma^2 + \gamma^3\right)$$

where $\gamma = Q_{\mathbf{y}'} + Q_{\mathbf{r}'_X} - Q_{\mathbf{x}'} + Q_{\mathbf{r}'_Y}$, and $\sigma = 1/\epsilon$.

The proof of these results is in the appendix. The first result says that the probability that we release an IQR score just because too much noise was added to A_j is small. The second result says that with high probability we recover the true causal direction, depending on the size of the dataset.

5 Training Set Privacy

Let $(\mathbf{x}, \mathbf{y})$ be the initial training data and $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ be the training data after a change in the dataset. Note that $\mathbf{x}$ and $\tilde{\mathbf{x}}$ differ in at most one element (similarly for $\mathbf{y}$ and $\tilde{\mathbf{y}}$). The length of both training datasets is n. From Algorithm 1, the only way the training set can affect the dependency scores $s_{X\rightarrow Y}, s_{Y\rightarrow X}$ is through the regression functions $\hat{f}, \hat{g}$, used to compute test set residuals $\mathbf{r}'_Y, \mathbf{r}'_X$. We use the kernel ridge regression method and so the functions $\hat{f}$ (and $\hat{g}$) can be written in the form: $\hat{f}(\mathbf{w}, x) = \mathbf{w}^\top \phi(x)$, where $\phi(x)$ is a (possibly infinite) feature space mapping to the Hilbert space corresponding to the kernel function used. Similar to other work on private regression [34] we assume that $|x|, |y| \leq 1$. The ridge regression algorithm can now be written as:

$$\mathbf{w} = \underset{\mathbf{w}\in\mathcal{H}}{\operatorname{argmin}} \frac{\lambda}{2}\|\mathbf{w}\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\phi(x_i) - y_i)^2, \quad (5)$$

where $\mathcal{H}$ is the corresponding Hilbert space. Practically speaking, even though $\mathbf{w}$ may be infinite-dimensional, because it always appears in an inner product with the feature mapping $\phi(x)$ we can utilize the 'kernel trick': $k(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$ to avoid having to represent $\mathbf{w}$ explicitly.

Let $\hat{f}(\mathbf{w}^*, \cdot)$ and $\hat{f}(\tilde{\mathbf{w}}^*, \cdot)$ be the classifiers resulting from the optimization problem in eq. (5) when trained on $(\mathbf{x}, \mathbf{y})$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, respectively (and similarly for $\hat{g}$). We show that the residuals in Algorithm 1 are bounded.

Theorem 5. *Say $\lambda \leq 1$. Given that the classifiers $\hat{f}(\mathbf{w}^*, \cdot), \hat{f}(\tilde{\mathbf{w}}^*, \cdot)$ are the result of the optimization problem in eq. (5), the residuals of these functions $\mathbf{r}'_Y, \tilde{\mathbf{r}}'_Y$ are bounded as,*

$$|r'_{i,Y} - \tilde{r}'_{i,Y}| \leq \frac{8}{n\lambda^{3/2}} \quad (6)$$

for all i, where $r'_{i,Y}, \tilde{r}'_{i,Y}$ are the ith elements of $\mathbf{r}'_Y, \tilde{\mathbf{r}}'_Y$ and m is the size of the test set.

This bound holds equally for $\mathbf{r}'_X, \tilde{\mathbf{r}}'_X$. The proof of the above is inspired by the work of Shalev-Shwartz et al. [31] and Jain & Thakurta [14]. We place the proof in the appendix for the interested reader. As far as we are aware this is the tightest bound for the optimization problem in eq. (5), with a non-Lipschitz loss. In the following, we use this bound to preserve training set privacy for the dependence scores considered in the previous section.

5.1 Rank Correlation Coefficients

Note that the bound in Theorem 5 directly implies that the ranking dependence scores have global sensitivity 1 (equal to the size of their ranges). To see this note that we can consider an adversarial situation in which the rank of every element of the residual $\mathbf{r}'_Y$ changes when the training set is altered in one element (as all the residual elements may change). This means that the Laplace mechanism cannot guarantee useful privacy.

Instead, note that both ranking scores may still have reasonably bounded local sensitivity. Specifically, if we consider the list of sorted residuals, it may be that there are large gaps between neighboring residuals. If this is the case then changing the training set by one point may not change the residual rankings. Thus, the

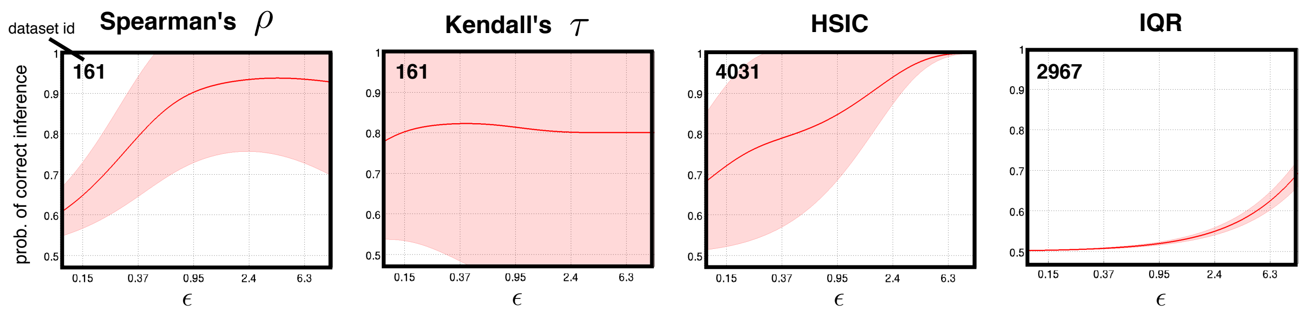

Figure 1: Probability of correctly identifying the causal direction on datasets selected from the Cause-Effect Pairs Challenge [12]. Datasets for which the scores perform well were selected in order to isolate the effect of privatization on the scores.

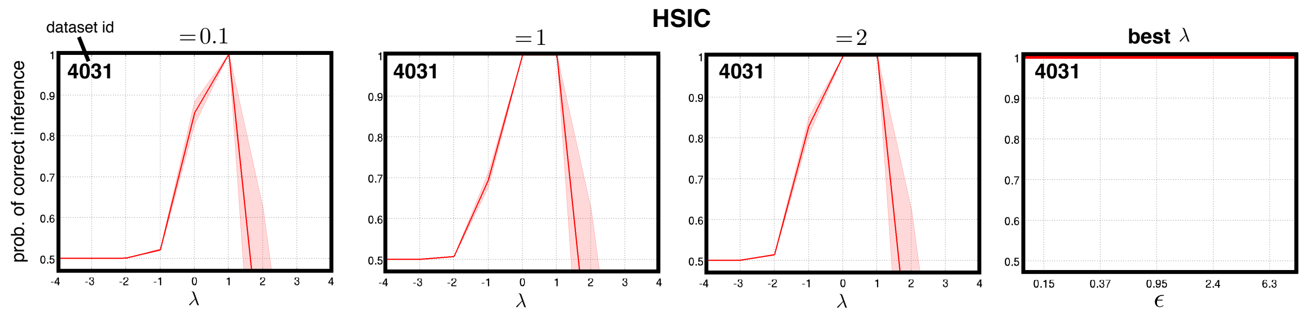

Figure 2: Training set privacy for the HSIC score. The three left-most plots show how λ affects the probability of correctly inferring the causal direction, while the right-most plot depicts this probability when the best λ is selected over a $\epsilon \in [0.1, 10]$. See text for more details.

ranking scores are in some sense stable to changes in the training set (for certain sets).

Definition 4. *We call a function f **k-stable on dataset** $\mathcal{D}$ if modifying any k elements in $\mathcal{D}$ does not change the value of f. Specifically, $f(\mathcal{D}) = f(\mathcal{D}^*)$ for all $\mathcal{D}^*$ such that $\mathcal{D}$ can be transformed into $\mathcal{D}^*$ with a minimum of k element substitutions. We say f is **unstable** on $\mathcal{D}$ if it is not even 1-stable on $\mathcal{D}$. The **distance to instability** of a dataset $\mathcal{D}$ w.r.t. a function f is the number of elements that must be changed to reach an unstable dataset.*

With these definitions, we will use a modification of the Propose-Test-Release framework that makes use of this stability as described in Algorithm 13 in Dwork & Roth [5].

Theorem 6. *[5] Algorithm 13 [5] is (ϵ, δ)-differentially private. Further, for all $\beta > 0$ if $s(\mathbf{x}', \mathbf{r}'_Y)$ is $\frac{\log(1/\delta)+\log(1/\beta)}{\epsilon}$-stable on $\mathbf{r}'_Y$, then Algorithm 13 releases $s(\mathbf{x}', \mathbf{r}'_Y)$ w.p. at least $1-\beta$.*

A lower bound on the distance to instability d is easily given by noting that $s(\mathbf{x}', \mathbf{r}'_Y)$ always outputs the same result as long as none of the ranks of $\mathbf{r}'_Y$ change. Let γ be the smallest absolute distance between any two ranks. Then a lower bound on d is, $d > \lfloor n\gamma\lambda^{3/2}/16 \rfloor$. This is the largest number of training points that may change so that the closest ranks moving towards each other do not overlap (given that they change by at most the amount in eq. 6). This lower-bound is sufficient to use Algorithm 13 [5] to privatize the ranking dependence scores.

5.2 HSIC Score

Theorem 7. *For $m \geq 2$, with kernels $k, l \leq 1$ where l is L_l-Lipschitz, the HSIC score has a training set sensitivity as follows,*

$$\left| \widehat{HSIC}_{k,l}(\mathbf{x}', \mathbf{r}'_Y) - \widehat{HSIC}_{k,l}(\mathbf{x}', \tilde{\mathbf{r}}'_Y) \right| \leq R\frac{32 L_l \sqrt{m}}{n}$$

where $R = \frac{8}{\lambda^{3/2}}$.

The proof follows directly from Theorem 5 and Lemma 16 in Mooij et al. [23]. Thus, the Laplace mechanism gives us $(\epsilon, 0)$-differential privacy and Theorem 2 gives us our utility guarantee.

5.3 IQR Score

Similar to the test set privacy section we will use propose-test-release to give a useful, private IQR score. In fact, we will use IQR algorithm almost identically, except that we will define A_j as the number of training points required to move the IQR out of an interval. Note that a lower bound on A_j is simply the number of points required to move every input less than the median to the left and every input larger than the median

Table 2: The non-private accuracies of the ANM model on a subset of the Cause-Effect Pairs Challenge [12], as well as the probability of correct causal inference after privatization.

DATASET IDS	4031	597	2209	2967	161	2132	1656	901	3484	1627
SIZE	7713	7748	7766	7771	7782	7784	7803	7820	7853	7862
	$\epsilon = \infty$ (NON-PRIVATE ACCURACIES)									
Spearman's ρ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.70 ± 0.48	0.90 ± 0.32	1.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.48	0.00 ± 0.00	1.00 ± 0.00
Kendall's τ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.70 ± 0.48	0.80 ± 0.42	1.00 ± 0.00	0.00 ± 0.00	0.80 ± 0.42	0.00 ± 0.00	1.00 ± 0.00
HSIC [11]	1.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.70 ± 0.48	0.60 ± 0.52	1.00 ± 0.00	0.40 ± 0.52	1.00 ± 0.00	0.10 ± 0.32
IQR [1]	0.50 ± 0.53	0.00 ± 0.00	0.10 ± 0.32	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.90 ± 0.32	0.00 ± 0.00	1.00 ± 0.00
	$\epsilon = 0.1$									
Spearman's ρ	0.56 ± 0.45	0.03 ± 0.00	0.20 ± 0.02	0.57 ± 0.10	0.61 ± 0.06	0.92 ± 0.02	0.40 ± 0.06	0.34 ± 0.21	0.01 ± 0.00	0.82 ± 0.02
Kendall's τ	0.54 ± 0.48	0.00 ± 0.00	0.00 ± 0.00	0.69 ± 0.38	0.78 ± 0.24	1.00 ± 0.00	0.12 ± 0.09	0.76 ± 0.41	0.00 ± 0.00	1.00 ± 0.00
HSIC [11]	0.68 ± 0.17	0.49 ± 0.00	0.60 ± 0.01	0.50 ± 0.00	0.50 ± 0.01	0.50 ± 0.00	0.52 ± 0.00	0.43 ± 0.06	0.66 ± 0.03	0.50 ± 0.00
IQR [1]	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.51 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
	$\epsilon = 1$									
Spearman's ρ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.69 ± 0.43	0.91 ± 0.17	1.00 ± 0.00	0.06 ± 0.07	0.30 ± 0.41	0.00 ± 0.00	1.00 ± 0.00
Kendall's τ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.70 ± 0.48	0.81 ± 0.40	1.00 ± 0.00	0.00 ± 0.00	0.80 ± 0.42	0.00 ± 0.00	1.00 ± 0.00
HSIC [11]	0.85 ± 0.16	0.39 ± 0.03	0.98 ± 0.00	0.52 ± 0.01	0.55 ± 0.06	0.50 ± 0.01	0.66 ± 0.02	0.21 ± 0.25	1.00 ± 0.01	0.49 ± 0.01
IQR [1]	0.54 ± 0.04	0.48 ± 0.00	0.49 ± 0.00	0.52 ± 0.00	0.58 ± 0.01	0.51 ± 0.01	0.48 ± 0.00	0.50 ± 0.00	0.47 ± 0.01	0.51 ± 0.00
	$\epsilon = 2$									
Spearman's ρ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.69 ± 0.47	0.93 ± 0.17	1.00 ± 0.00	0.01 ± 0.02	0.31 ± 0.45	0.00 ± 0.00	1.00 ± 0.00
Kendall's τ	0.50 ± 0.53	0.00 ± 0.00	0.00 ± 0.00	0.70 ± 0.48	0.80 ± 0.42	1.00 ± 0.00	0.00 ± 0.00	0.80 ± 0.42	0.00 ± 0.00	1.00 ± 0.00
HSIC [11]	0.92 ± 0.09	0.29 ± 0.04	1.00 ± 0.00	0.55 ± 0.01	0.59 ± 0.11	0.51 ± 0.01	0.78 ± 0.02	0.20 ± 0.26	1.00 ± 0.00	0.48 ± 0.02
IQR [1]	0.58 ± 0.09	0.46 ± 0.01	0.49 ± 0.01	0.54 ± 0.01	0.65 ± 0.02	0.52 ± 0.02	0.47 ± 0.01	0.51 ± 0.01	0.45 ± 0.01	0.52 ± 0.01

to the right (or the reverse of these), using the bound on $\mathbf{r}$ in eq. (6). The aforementioned privacy and utility results of the IQR propose-test-release framework apply here. The only difference is we just need to add noise to the IQR scores computed on the residuals, which implies $(6\epsilon, 2\delta)$-privacy and that the results of Theorem 4 can be tightened.

6 Results

We test our methods for private release of causal inference statistics on a small subsets from the Cause-Effect Pairs Competition collection Guyon [12]. Specifically, we randomly select 10 of the largest 25 datasets that have a causal direction either $X \to Y$ or $Y \to X$. We average over 10 random 50/50 train/test splits of the data. Table 2 shows the non-private accuracy of the four dependence scores over these datasets. We show the probability of correct causal inference changes as these scores are made private w.r.t. the test set. Note that these scores are often complementary, with the ranking-based scores performing well on datasets in which HSIC does worse, and vice-versa.

Figure 1 shows the effect of privatization on the dependence scores: HSIC and IQR. Note that, for low ϵ (increased privacy), the probability of correct influence is lower as the amount of noise required blurs the true dependence scores. However, as ϵ increases, so does this probability, in some cases drastically. For the IQR score, recall that there is a probability that the algorithm returns null: $\perp$, if R_j is less than a threshold controlled by δ. We investigated this probability, by varying $\delta \in [10^{-5}, 10^{-2}]$ and sampling $10,000$ points from the appropriate Laplace distribution. We found that, for the IQR dataset in figure 1 *every sample* did not move R_j below the null threshold. Therefore, the probability of null is essentially 0.

The three left-most plots in Figure 2 demonstrate how λ, which has a large effect on the training set sensitivity (as described in eq. 6) affects the probability of correct inference. We perform this experiment for different settings of ϵ, and each one produces a distinctive 'hump' shape. This is because for small λ the sensitivity bound (6) is too large to produce meaningful causal inference. Similarly, for large λ the kernelized regression algorithm (5) is overly-regularized, which produces a poor regressor and poor dependence scores. Only when λ is within a certain range do we balance the size of the sensitivity bound with the size of the regularization. This range grows larger as ϵ increases as the privacy setting becomes less strict (requiring less noise). The right-most plot shows the correct inference probability using the best λ for a range of $\epsilon \in [0.1, 10]$. With proper selection of λ we can achieve high-quality causal inference that maintains privacy w.r.t. the training set.

7 Conclusion

We have presented, to the best of our knowledge, the first work towards differentially private causal inference. There are numerous directions of future work including privatizing other causal inference frameworks (e.g. IGCI [16]), analyzing that ANM algorithm without train/test splits, as well as other dependence scores. As there is significant overlap in the applications of causal inference and private learning we believe this work constitutes an important step towards making causal inference practical.

Acknowledgments

KQW and MJK are supported by NSF grants IIA-1355406, IIS-1149882, EFRI-1137211. We thank the anonymous reviewers for their useful comments.

References

[1] Bühlmann, Peter, Peters, Jonas, Ernest, Jan, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[2] Chaudhuri, Kamalika, Monteleoni, Claire, and Sarwate, Anand D. Differentially private empirical risk minimization. *JMLR*, 12:1069–1109, 2011.

[3] Dinur, Irit and Nissim, Kobbi. Revealing information while preserving privacy. In *Proceedings of the SIGMOD-SIGACT-SIGART symposium on principles of database systems*, pp. 202–210. ACM, 2003.

[4] Dwork, Cynthia and Lei, Jing. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

[5] Dwork, Cynthia and Roth, Aaron. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[6] Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–284. Springer, 2006.

[7] for Disease Control, Centers, Prevention, et al. *How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general*. Centers for Disease Control and Prevention (US), 2010.

[8] Friedman, Nir and Nachman, Iftach. Gaussian process networks. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 211–219. Morgan Kaufmann Publishers Inc., 2000.

[9] Geiger, Philipp, Janzing, Dominik, and Schölkopf, Bernhard. Estimating causal effects by bounding confounding. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 240–249, 2014.

[10] Geiger, Philipp, Zhang, Kun, Schoelkopf, Bernhard, Gong, Mingming, and Janzing, Dominik. Causal inference by identification of vector autoregressive processes with hidden components. In *ICML*, pp. 1917–1925, 2015.

[11] Gretton, Arthur, Bousquet, Olivier, Smola, Alex, and Schölkopf, Bernhard. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pp. 63–77. Springer, 2005.

[12] Guyon, I. Cause-effect pairs kaggle competition, 2013. URL `https://www.kaggle.com/c/cause-effect-pairs/`.

[13] Hoyer, Patrik O, Janzing, Dominik, Mooij, Joris M, Peters, Jonas, and Schölkopf, Bernhard. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

[14] Jain, Prateek and Thakurta, Abhradeep. Differentially private learning with kernels. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 118–126, 2013.

[15] Jain, Prateek, Kothari, Pravesh, and Thakurta, Abhradeep. Differentially private online learning. *COLT*, 2012.

[16] Janzing, Dominik, Mooij, Joris, Zhang, Kun, Lemeire, Jan, Zscheischler, Jakob, Daniušis, Povilas, Steudel, Bastian, and Schölkopf, Bernhard. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[17] Kano, Yutaka and Shimizu, Shohei. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pp. 261–270, 2003.

[18] Kpotufe, Samory, Sgouritsa, Eleni, Janzing, Dominik, and Schölkopf, Bernhard. Consistency of causal inference under the additive noise model. In *ICML*, 2014.

[19] Lopez-Paz, David, Muandet, Krikamol, Schölkopf, Bernhard, and Tolstikhin, Iliya. Towards a learning theory of cause-effect inference. In *ICML*, 2015.

[20] McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. In *FOCS*, pp. 94–103. IEEE, 2007.

[21] Mooij, Joris M, Stegle, Oliver, Janzing, Dominik, Zhang, Kun, and Schölkopf, Bernhard. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pp. 1687–1695, 2010.

[22] Mooij, Joris M, Janzing, Dominik, Heskes, Tom, and Schölkopf, Bernhard. On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, pp. 639–647, 2011.

[23] Mooij, Joris M, Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob, and Schölkopf, Bernhard. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

[24] Nissim, Kobbi, Raskhodnikova, Sofya, and Smith, Adam. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84. ACM, 2007.

[25] Pearl, Judea. Causality: models, reasoning, and inference. 2000.

[26] Peters, Jonas, Mooij, Joris M, Janzing, Dominik, and Schölkopf, Bernhard. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[27] Rasmussen, Carl Edward and Williams, Christopher K. I. Gaussian processes for machine learning. 2006.

[28] Reichenbach, Hans and Reichenbach, Maria. *The direction of time*. Univ of California Press, 1956.

[29] Sgouritsa, Eleni, Janzing, Dominik, Hennig, Philipp, and Schölkopf, Bernhard. Inference of cause and effect with unsupervised inverse regression. In *AISTATS*, pp. 847–855, 2015.

[30] Shajarisales, Naji, Janzing, Dominik, Shoelkopf, Bernhard, and Besserve, Michel. Telling cause from effect in deterministic linear dynamical systems. In *ICML*, 2015.

[31] Shalev-Shwartz, Shai, Shamir, Ohad, Srebro, Nathan, and Sridharan, Karthik. Stochastic convex optimization. In *COLT*, 2009.

[32] Spirtes, Peter, Glymour, Clark N, and Scheines, Richard. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[33] Sun, Xiaohai, Janzing, Dominik, and Schölkopf, Bernhard. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, 71(7):1248–1256, 2008.

[34] Talwar, Kunal, Thakurta, Abhradeep, and Zhang, Li. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

[35] Zhang, Kun and Hyvärinen, Aapo. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655. AUAI Press, 2009.