Consistency of Causal Inference under the Additive Noise Model

Samory Kpotufe SAMORY@TTIC.EDU
Toyota Technological Institute-Chicago

Eleni Sgouritsa ELENI.SGOURITSA@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems

Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems

Abstract

We analyze a family of methods for statistical causal inference from sample under the so-called *Additive Noise Model*. While most work on the subject has concentrated on establishing the soundness of the Additive Noise Model, the statistical consistency of the resulting inference methods has received little attention. We derive general conditions under which the given family of inference methods consistently infers the causal direction in a nonparametric setting.

1. Introduction

Drawing causal conclusions for a set of observed variables given a sample from their joint distribution is a fundamental problem in science. Conditional-independence-based methods (Pearl, 2000; Spirtes et al., 2000) estimate a set of directed acyclic graphs, all entailing the same conditional independences, from the data. However, these methods can not distinguish between two graphs that entail the same set of conditional independences, the so-called *Markov equivalent* graphs. Consider for example the case of only two observed dependent random variables. Conditional-independence-based methods can not recover the causal graph since $X \to Y$ and $Y \to X$ are Markov equivalent. An elegant basis for causal graphs is the framework of structural causal models (SCMs) (Pearl, 2000), where every observable is a function of its parents and an unobserved independent noise term. This allows us to formulate an assumption on function classes which lets us infer the causal direction in two-variable case.

A special case of SCMs is the *Causal Additive Noise Model* (CAM) (Shimizu et al., 2006; Hoyer et al., 2009; Tillman et al., 2009; Peters et al., 2011a;b) which is given as follows: given two random variables X and Y, X is assumed to cause Y if (i) Y can be obtained as a function of X plus a noise term independent of X, but (ii) X cannot be obtained as a function of Y plus independent noise, then we infer that X causes Y. In this case, where (i) and (ii) hold simultaneously, the CAM is termed *identifiable*.

Initial work on the CAM focused on establishing its theoretical soundness, i.e. understanding the class of distributions $P_{X,Y}$ for which the CAM is identifiable, i.e. for which (i) and (ii) hold simultaneously. Early work by (Shimizu et al., 2006) showed that the CAM is identifiable when the functional relationship $Y = f(X) + \eta$ is linear, provided the independent noise η is not Gaussian. Later, Hoyer et al. (2009), Zhang & Hyvärinen (2009) and Peters et al. (2011a) showed that the CAM is identifiable more generally even if f is nonlinear, the main technical requirements being that the marginals P_X, and P_η are absolutely continuous on $\mathbb{R}$, with P_η having support $\mathbb{R}$. Note that Zhang & Hyvärinen (2009) also introduces a generalization of the CAM termed post-nonlinear models. Further work by Peters et al. (2011b) showed how to reduce causal inference for a network of multiple variables under the CAM to the case of two variables X and Y discussed so far, by properly extending the conditions (i) and (ii) to conditional distributions instead of marginals. Thus, the soundness of the CAM being established by these various works, the next natural question is to understand the statistical behavior of the resulting estimation procedures on finite samples.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Current insights into this last question are mostly empirical. Various works (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011a) have successfully validated procedures based on the CAM (outlined in Section 1.1 below) on a mix of artificial and real-world datasets where the causal structure to be inferred is clear. However, on the theoretical side, it remains unclear whether these procedures can infer causality from samples in general situations where the CAM is identifiable. In the particular case where the functional relation between X and Y is linear, Hyvärinen et al. (2008) proposed a successful method shown to be consistent. Two recent Arxived results, Bühlmann et al. (2013); Nowzohour & Bühlmann (2013), show the consistency of maximum log-likelihood approaches to causal inference under the multi-variable network extension of Peters et al. (2011b).

While consistency has been shown for particular procedures, in this paper we are rather interested in general conditions under which common approaches, with various algorithmic instantiations, are consistent. We derive both algorithmic and distributional conditions for statistical consistency in general situations where the CAM is identifiable. The present work focuses on the case of two real variables, allowing us to focus on the inherent difficulties of achieving consistency with the common algorithmic approaches. These difficulties, described in Section 1.2 have to do with estimating the *degree* of independence between noise and input, while the noise is itself estimated from the input and hence is inherently dependent on the input.

1.1. Inference Methods Under the Additive Noise Model

Causal inference methods under the Additive Noise Model typically follow the meta-procedure below. Assume f and g are the best functional fits under some risk, respectively $Y \approx f(X)$ and $X \approx g(Y)$:

> Fit Y as a function $f(X)$, obtain the residuals $\eta_{Y,f} = Y - f(X)$, fit X as a function $g(Y)$, obtain the residuals $\eta_{X,g} = X - g(Y)$, decide $X \to Y$ if $\eta_{Y,f} \perp\!\!\!\perp X$ but $\eta_{X,g} \not\perp\!\!\!\perp Y$, decide $Y \to X$ if the reverse holds true, abstain otherwise.

Instantiations thus vary in the regression procedures employed for function fitting, and in the independence measures employed. Our analysis concerns procedures employing an entropy-based independence measure, which is cheaper than usual independence tests. These procedures vary in the regression and entropy estimators employed. They are presented in detail in Section 3.

1.2. Towards Consistency: Main Difficulties

Assume (i) and (ii) hold so that X *causes* Y under the CAM. We want to detect this from sufficiently large finite samples. This is consistency in a rough sense.

Establishing consistency of the above meta-procedure faces many subtle difficulties. The above outlined algorithmic approach consists of four interdependent statistical estimation tasks, namely two regression problems and two independence-tests. Considered separately, the consistency of such estimation tasks is well understood, but in the present context the success of the independence tests is contingent on successful regression.

The main difficulty is that although we are observing X and Y, we are not observing the residuals $\eta_{Y,f}$ and $\eta_{X,g}$, but empirical approximations η_{Y,f_n} and η_{X,g_n} obtained by estimating f and g as f_n and g_n on a sample of size n.

For now, consider just detecting that $\eta_{Y,f}$, f unknown, is independent from X. A good estimator f_n will ensure that f_n and f are *close*, usually in an L_2 sense (i.e. $\mathbb{E}_X |f_n(X) - f(X)|^2 \approx 0$). Hence η_{Y,f_n} is *close* to $\eta_{Y,f}$, but unfortunately this does not imply that $\eta_{Y,f_n} \perp\!\!\!\perp X$ if $\eta_{Y,f} \perp\!\!\!\perp X$. In fact it is easy to construct r.v.'s A, B, C such that $A \perp\!\!\!\perp B$, $|B - C| < \epsilon$, for arbitrary ϵ, but $C \not\perp\!\!\!\perp A$. Thus, the estimate η_{Y,f_n} might be *close* to $\eta_{Y,f}$, yet it might still appear dependent on X even if $\eta_{Y,f}$ is not. Complicating matters further, η_{Y,f_n} and $\eta_{Y,f}$ would only be close in an average sense (instead of close for every value of X) since f_n and f are typically only close in an average sense (e.g. close in L_2).

Now consider the full causal discovery, i.e. consider also detecting that $\eta_{X,g}$ depends on Y. To achieve consistency, the independence test employed must detect more dependence between η_{X,g_n} and Y than between η_{Y,f_n} and X. This will depend on how the particular independence test is influenced by errors in the particular regression procedures employed, and the relative rates at which these various procedures converge.

As previously mentioned, we will consider a family of independence-tests based on comparing sums of entropies. We will handle the above difficulties and derive conditions for consistency by first understanding how the various estimated entropies converge as a function of regression convergence (L_2 convergence).

We do not consider the question of finite-sample convergence rates for causal estimation under the CAM. In fact, it is not even clear whether it is generally possible to establish such rates. This is because it is generally possible that the Bayes best fits $f(x) = \mathbb{E}[Y|x]$ is smooth while $g(y) = \mathbb{E}[X|y]$ is not even continuous; yet it is well known that without smoothness or similar structural conditions, ar-

bitrarily bad rates of convergence are possible in regression (see e.g. (Gyorfi et al., 2002), Theorem 3.1).

However, along the way of deriving consistency, we analyze the convergence of various quantities, which appear to affect the finite-sample behavior of the meta-procedure. In particular the tails of the additive noise and the richness of the regression algorithms seem to have a strong effect on convergence. This is verified in controlled simulations. The theoretical details are discussed in Section 4.

2. Preliminaries

2.1. Setup and Notation

We let H and I denote respectively differential entropy, and mutual information (Cover et al., 1994). Given a density p we will at times use the (abuse of) notation $H(p)$ when a r.v. is unspecified.

The distribution of a r.v. Z is denoted P_Z, and its density when it exists is denoted p_Z.

Throughout the analysis we will be concerned with residuals from regression fits. We use the following notation.

Definition 1. *For a function $f : \mathbb{R} \mapsto \mathbb{R}$, we consider either of the* **residuals***: $\eta_{Y,f} \triangleq Y - f(X)$ and $\eta_{X,f} \triangleq X - f(Y)$.*

The Causal Additive Noise Model is captured as follows:

Definition 2 (CAM)**.** *Given r.v.'s X, Y, a function $f : \mathbb{R} \mapsto \mathbb{R}$ and a r.v. η, we write $X \xrightarrow{f,\eta} Y$ if the following holds:*

(i) *$P_{X,Y}$ is generated as $X \sim P_X$, and $Y = f(X) + \eta$, where the noise r.v. η has 0 mean and $\eta \perp\!\!\!\perp X$;*

(ii) *for any $g : \mathbb{R} \mapsto \mathbb{R}$, $\eta_{X,g} \triangleq X - g(Y)$ depends on X.*

We write $X \to Y$ when f and η are clear from context.

3. Causal Inference Procedures

3.1. Main Intuition

Lemma 1. *Consider any absolutely continuous joint-distribution $P_{X,Y}$ on $X, Y \in \mathbb{R}$. For any two functions $f, g : \mathbb{R} \mapsto \mathbb{R}$ we have*

$$H(X) + H(\eta_{Y,f}) = H(Y) + H(\eta_{X,g}) - \{I(\eta_{X,g}, Y) - I(\eta_{Y,f}, X)\}.$$

Proof. By the chain rule of differential entropy we have

$$H(X,Y) = H(X) + H(Y|X) = H(X) + H(\eta_{Y,f}|X) = H(X) + H(\eta_{Y,f}) - I(\eta_{Y,f}, X), \text{ similarly}$$

$$H(X,Y) = H(Y) + H(\eta_{X,g}) - I(\eta_{X,g}, Y).$$

Equate the two r.h.s above and rearrange. □

Note that whenever $\eta_{Y,f} \perp\!\!\!\perp X$, we have $I(\eta_{Y,f}, X) = 0$. Therefore, by the above lemma, if $\eta_{Y,f} \perp\!\!\!\perp X$ then $C_{XY} \triangleq H(X) + H(\eta_{Y,f})$ is smaller than $C_{YX} \triangleq H(Y) + H(\eta_{X,g})$. This yields a measure of independence which is relatively cheap to estimate. In particular the test depends only on the marginal distributions of the r.v.'s X, Y and functional residuals, and does not involve estimating joint distributions or conditionals, as is implicit in most independence tests. We analyze a family of procedures based on this idea. This family is given in the next subsection.

3.2. Meta-Algorithm

Let $\{(X_i, Y_i)\}_1^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a finite sample drawn from $P_{X,Y}$. Let $H_n(X)$ and $H_n(Y)$ be respective estimators of $H(X)$ and $H(Y)$ based on the sample $\{(X_i, Y_i)\}_1^n$.

We consider the following family of inference procedures:

> Given an i.i.d sample $\{(X_i, Y_i)\}_1^n$ from $P_{X,Y}$, let f_n be returned by an algorithm which fits Y as $f_n(X)$ and g_n be returned by an algorithm which fits X as $g_n(Y)$. Let H_n denote an entropy estimator. Given a threshold parameter $\tau_n \xrightarrow{n\to\infty} 0$:
>
> Decide $X \to Y$ if
>
> $$H_n(X) + H_n(\eta_{Y,f_n}) + \tau_n \leq H_n(Y) + H_n(\eta_{X,g_n}).$$
>
> Decide $Y \to X$ if
>
> $$H_n(Y) + H_n(\eta_{X,g_n}) + \tau_n \leq H_n(X) + H_n(\eta_{Y,f_n}).$$
>
> Abstain otherwise.

The analysis in this paper is carried with respect to the L_{2,P_X} and L_{2,P_Y} functional norms defined as follows.

Definition 3. *For $f : \mathbb{R} \mapsto \mathbb{R}$, and a measure μ on $\mathbb{R}$, the $L_{2,\mu}$ norm is given as $\|f\|_{2,\mu} = \left(\int_t f(t)^2 \, d\mu(t)\right)^{1/2}$.*

We assume the internal procedures f_n, g_n, H_n have the following consistency properties.

Assumption 1. *The internal procedures are consistent:*

- *Suppose $\mathbb{E}\, Y^2 < \infty$. Let $f(x) \triangleq \mathbb{E}[Y|x]$. Then $\|f_n - f\|_{2,P_X} \xrightarrow{P} 0$.*
- *Suppose $\mathbb{E}\, X^2 < \infty$. Let $g(y) \triangleq \mathbb{E}[X|y]$. Then $\|g_n - g\|_{2,P_Y} \xrightarrow{P} 0$.*
- *Suppose Z has bounded variance, and has continuous density p_Z such that $\exists\, T, C > 0, \alpha > 1, \quad \forall\, |t| > T$, $p_Z(t) \leq C\,|t|^{-\alpha}$. Then $|H_n(Z) - H(Z)| \xrightarrow{P} 0$.*

Many common nonparametric regression procedures (e.g. kernel, k-NN, Kernel-SVM, spline regressors) are consistent in the above sense (Gyorfi et al., 2002). Also the consistency of a variety of entropy estimators (e.g. plug-in entropy estimators) is well established (Beirlant et al., 1997).

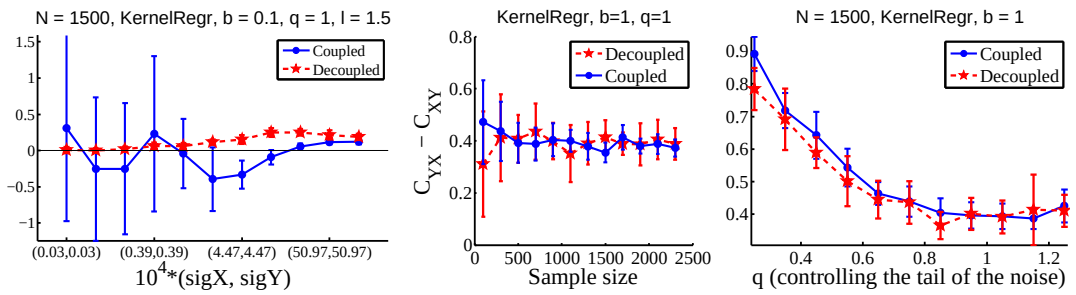

Figure 1. Plots of the difference between the complexity measures ($C_{YX} - C_{XY}$) for coupled and decoupled-estimation in various scenarios. Simulated data is generated as $Y = bX^3 + X + \eta$. X is sampled from a uniform distribution on the open interval $(-2.5, 2.5)$, while η is sampled as $|\mathcal{N}|^q \cdot \text{sign}(\mathcal{N})$ where $\mathcal{N}$ is a standard normal. b controls the strength of the nonlinearity of the function and q controls the non-Gaussianity of the noise: $q = 1$ gives a Gaussian, while $q > 1$ and $q < 1$ produces super-Gaussian and sub-Gaussian distributions, respectively. For entropy estimation we employ a resubstitution estimate using a kernel density estimator tuned against log-likelihood (Beirlant et al., 1997) and for regression estimator we use kernel regression (KR). For every combination of the parameters, each experiment was repeated 10 times, and average results for ($C_{YX} - C_{XY}$) are reported along with standard deviation across repetitions. Plot (a): increasing kernel bandwidth of regressor geometrically (by factors of $l = 1.5$), i.e. decreasing richness of the algorithm. When the capacity of the regression algorithm is too large, the variance of the causal inference is large for coupled-estimation (due to overfitting) but remains low for decoupled-estimation. Plot (b): increasing sample size (bandwidth of KR tuned by cross-validation). For tuned bandwidth, the variance of the causal inference is only due to the sample size, so the coupled-estimation (which estimates everything on a larger sample) becomes the better procedure. Plot (c): increasing q, i.e. the tail of the noise is made sharper (KR tuned by cross-validation). For faster decreasing tail of the noise, the causal inference becomes better. The experiments of Figures (b) and (c) were repeated using kernel ridge regression (KRR) tuned by cross-validation (see appendix of the extended paper). For properly tuned parameters, the selection of regression method does not seem to matter for the causal inference results.

4. Technical overview of results

We consider the following two versions of the above meta-procedure. The analysis (Section 5) is divided accordingly.

Definition 4 (Decoupled-estimation)**.** *f_n and g_n are learned on half of the sample $\{(X_i, Y_i)\}_1^n$, and the $H_n(\eta_{Y,f_n})$ and $H_n(\eta_{X,g_n})$ are learned on the other half of the sample (w.l.o.g. assume n is even). $H_n(X)$ and $H_n(Y)$ could be learned on either half or on the entire sample.*

Definition 5 (Coupled-estimation)**.** *All f_n, g_n and entropies H_n are learned on the entire sample $\{(X_i, Y_i)\}_1^n$.*

Our most general consistency result (Theorem 1, Section 5.1) concerns decoupled-estimation. By decoupling regression and entropy estimations, we reduce the potential of overfitting, during entropy estimation, the generalization error of regression. This generalization error could be large if the regression algorithms are too rich (e.g. ERM over large functional classes). Our simulations show that, when the regression algorithm is too rich, the variance of the causal inference is large for coupled-estimation but remains low for decoupled-estimation (Fig. 1(a)). By decreasing the richness of the class (simulated by increasing the kernel bandwidth for a kernel regressor) the source of variance shifts to the sample size, and coupled-estimation (which estimates everything on a larger sample) becomes the better procedure and tends to converge faster (Fig. 1(b)).

For the consistency result of Theorem 1 we make no assumption on the richness of the regression algorithms, but simply assume that they converge in L_2 (Assumption 1). The main technicality is to then show that entropies of residuals are locally continuous relative to the L_2 metric in both causal and anticausal directions.

For coupled-estimation, the main difficulty is the following. Even though the entropy estimators are consistent for a *fixed* distribution, the distribution of the residuals change with f_n and g_n, thus with every random sample (this problem is alleviated by decoupling the estimation). However, if the richness of the regression algorithms is controlled, in other words if the set of potential f_n and g_n is not too rich, then the entropy estimate for residuals might converge. We show in Theorem 2 (Section 5.2) that if we employ kernel regressors with properly chosen bandwidths, and kernel-based entropy estimators with sufficiently smooth kernels, then the resulting method is consistent for causal inference.

Both consistency results of Theorem 1 and Theorem 2 rely on tail assumptions on the additive noise η (where $X \xrightarrow{f,\eta} Y$). We assume an exponentially decreasing tail for the more difficult case of coupled-estimation, but need only a mild assumption of polynomially decreasing tail in the case of decoupled-estimation. Note that it is common to assume that η has Gaussian tail, and our assumptions are milder in that respect.

Interestingly, our analysis for Theorem 1 suggests that convergence of causal inference is likely faster if the noise η has faster decreasing tail (see Lemma 3). This is verified in our simulations where we vary the tail of η (Fig. 1(c)).

5. Analysis

5.1. Consistency for Decoupled-estimation

In this section we establish a general consistency result for the meta-procedure above. The main technicality consists of relating differential entropy of residuals to the L_2-norms of residuals (i.e. to the error made in function estimation). We henceforth let Σ denote the Lebesgue measure.

The analysis in this section uses the following polynomial tail assumption on η. We note that Assumption 2 satisfies the idenfiability conditions of (Zhang & Hyvärinen, 2009).

Assumption 2 (Tail)**.** *$P_{X,Y}$ is generated as follows: $X \xrightarrow{f,\eta} Y$ for some bounded function f, with bounded derivative on $\mathbb{R}$. P_X has bounded support, and both P_X and P_η have densities p_X, p_η with bounded derivatives on $\mathbb{R}$. Furthermore, we assume η has bounded variance, and p_η satisfies, for some $T > 0$, $C > 0$, and $\alpha > 1$:*

$$\forall\, |t| > T, \quad p_\eta(t) \leq C\, |t|^{-\alpha}\,, \tag{1}$$

Note that, since the unknown target functions are assumed bounded, any consistent regressor can be appropriately truncated while maintaining consistency. We therefore have the following technical assumption on the regressors.

Assumption 3. *The regression procedures return bounded functions:* $\lim_{n\to\infty} \max\left\{\|f_n(t)\|_\infty\,, \|g_n(t)\|_\infty\right\} < \infty$.

Theorem 1 (General consistency for decoupled-estimation)**.** *Suppose $X \xrightarrow{f,\eta} Y$ for some f, η, and $P_{X,Y}$ satisfies the tail Assumption 2. Suppose f_n, g_n, and H_n are consistent procedures satisfying Assumption 1 and 3. Let the meta-algorithm be decoupled as in Definition 4.*

Then the probability of correctly deciding $X \to Y$ goes to 1 as $n \to \infty$.

To prove the theorem, we have to understand how the estimated entropies converge as a function of the L_2 error in regression estimation. We will proceed by bounding the distance between the densities $p_{\eta_{Y,f}}$ and $p_{\eta_{Y,f'}}$ of the residuals of functions f and f' in terms of the L_2 distance between f and f' (Lemma 3); this will then be used to bound the difference in the entropy of such residuals.

Given Assumption 2, the following lemma establishes some useful properties of the distribution $P_{X,Y}$ and of the distribution of certain residuals. It is easy to verify that under our assumptions, all distributions under consideration in the lemma are absolutely continuous.

Lemma 2 (Properties of induced densities)**.** *Suppose $P_{X,Y}$ satisfies Assumption 2 for some f, η, and $\alpha > 1$. We then have the following: (*i*) $p_{X,Y}$ has a bounded gradient on $\mathbb{R}^2$, (*ii*) consider functions $f', g : \mathbb{R} \mapsto \mathbb{R}$ and suppose $\sup|f'|$ and $\sup|g|$ are at most T_0 for some T_0; then there exists $T' > 0$ depending on T_0, and $C' > 0$ such that $\forall\, |t| > T'$*

$$\left\{p_{X,Y}(\cdot, t), p_{X,Y}(t, \cdot), p_{\eta_{Y,f'}}(t), p_{\eta_{X,g}}(t)\right\} \leq C'\, |t|^{-\alpha}\,.$$

In particular, the above holds for $g(y) \triangleq \mathbb{E}\left[X|Y = y\right]$.

The next lemma relates the density of residuals to the L_2 distance between functions. Notice, as discussed in Section 4, that the Lemma suggests that the densities of residuals converge faster the sharper the tails of the noise η: the larger α, the sharper the bounds are in terms of the L_2 distance between functions.

Lemma 3 (Density of residuals w.r.t. L_2 distance)**.** *Suppose the joint distribution $P_{X,Y}$ satisfies Assumption 2 for some f, η and $\alpha > 1$. Let $g(y) \triangleq \mathbb{E}[X|Y = y]$. Consider functions $f', g' : \mathbb{R} \mapsto \mathbb{R}$. There exist a constant C'' such that for $\|f - f'\|_{2,P_X}$ and (respectively) $\|g - g'\|_{2,P_Y}$ sufficiently small,*

$$\sup_{t\in\mathbb{R}} \left|p_{\eta_{Y,f'}}(t) - p_{\eta_{Y,f}}(t)\right| \leq C'' \left(\|f' - f\|_{2,P_X}\right)^{(\alpha-1)/2\alpha},$$

$$\sup_{t\in\mathbb{R}} \left|p_{\eta_{X,g'}}(t) - p_{\eta_{X,g}}(t)\right| \leq C'' \left(\|g' - g\|_{2,P_Y}\right)^{(\alpha-1)/2\alpha}.$$

Proof. We start by bounding the difference between $p_{\eta_{Y,f'}}(t)$ and $p_{\eta_{Y,f'}}(t)$. We note that the same ideas can be used to bound the difference between $p_{\eta_{X,g'}}(t)$ and $p_{\eta_{X,g}}(t)$, since X and Y are interchangeable in the analysis from this point on. This is because what follows does not depend on how $P_{X,Y}$ is generated, just on the properties of the induced distributions as stated in Lemma 2.

We will partition the space $\mathbb{R}$ as follows. First, let $\mathbb{R}_>$ denote the set $\left\{x : |f(x) - f'(x)| > \sqrt{\|f - f'\|_{2,P_X}}\right\}$. We define the following interval $\mathcal{U} \subset \mathbb{R}$: let T' be defined as in Lemma 2, and $\tau > T'$; we have $\mathcal{U} \triangleq [-\tau, \tau]$.

For any $t \in \mathbb{R}$ we have by writing residual densities in terms of the joint $p_{X,Y}$ (as in the proof of Lemma 2 in the

appendix of the extedned version) that $\left|p_{\eta_{Y,f'}}(t) - p_\eta(t)\right|$

$$= \left|\int_{\mathbb{R}} \left(p_{X,Y}(x, t + f'(x)) - p_{X,Y}(x, t + f(x))\right)\, dx\right|$$

$$\leq \left|\int_{\mathbb{R}\setminus D} \left(p_{X,Y}(x, t + f'(x)) - p_{X,Y}(x, t + f(x))\right)\, dx\right| \tag{2}$$

$$+ \int_{\mathcal{U}\setminus\mathbb{R}_>} \left|p_{X,Y}(x, t + f'(x)) - p_{X,Y}(x, t + f(x))\right|\, dx \tag{3}$$

$$+ \left|\int_{\mathbb{R}_>} \left(p_{X,Y}(x, t + f'(x)) - p_{X,Y}(x, t + f(x))\right)\, dx\right|. \tag{4}$$

To bound the first term (2), let y_x denote either of $t + f'(x)$ or $t + f(x)$, we have by Lemma 2 that

$$\int_\tau^\infty p_{X,Y}(x, y_x)\, dx \leq \int_\tau^\infty C' x^{-\alpha}\, dx \leq \frac{C'}{\alpha - 1}\tau^{-(\alpha-1)},$$

so that the first term (2) is at most $2\frac{C'}{\alpha-1}\tau^{-(\alpha-1)}$.

To bound the second term (3) we recall that $p_{X,Y}$ has a bounded gradient on $\mathbb{R}^2$ (Lemma 2). Therefore there exists C_0 such that for every $x, y, \epsilon \in \mathbb{R}$, $p_{X,Y}(x, y + \epsilon)$ differs from $p_{X,Y}(x, y)$ by at most $C_0 \cdot |\epsilon|$. It follows that the second term (3) is at most

$$\int_{\mathcal{U}\setminus\mathbb{R}_>} C_0 \left|f'(x) - f(x)\right|\, dx \leq 2\tau \cdot C_0 \sqrt{\|f - f'\|_{2,P_X}}.$$

The third term (4) is equal to

$$\left|\mathbb{P}\left(X \in \mathbb{R}_>, Y = t + f'(X)\right) - \mathbb{P}\left(X \in \mathbb{R}_>, Y = t + f(X)\right)\right| \leq P_X(\mathbb{R}_>).$$

We next bound $P_X\left(\mathbb{R}_>\right)$ while noting that $\|f - f'\|_{2,P_X}$ could be 0. Let $\epsilon > \|f - f'\|_{2,P_X}$. By Markov's inequality,

$$P_X\left\{|f(X) - f'(X)| > \sqrt{\epsilon}\right\} \leq \frac{\|f - f'\|_{1,P_X}}{\sqrt{\epsilon}} \leq \frac{\|f - f'\|_{2,P_X}}{\sqrt{\epsilon}}.$$

Thus, consider a sequence of $\epsilon \to \|f - f'\|_{2,P_X}$, by Fatou's lemma we have $P_X\left(\mathbb{R}_>\right) \leq \sqrt{\|f - f'\|_{2,P_X}}$.

Combining the above analysis we have that

$$\left|p_{\eta_{Y,f'}}(t) - p_\eta(t)\right| \leq 2\frac{C'}{\alpha - 1}\tau^{-(\alpha-1)} + (1 + 2\tau \cdot C_0)\sqrt{\|f - f'\|_{2,P_X}}.$$

Now, for $\|f - f'\|_{2,P_X}$ sufficiently small, we can pick $\tau = O\left(\|f - f'\|_{2,P_X}\right)^{-1/2\alpha}$ to get the result.

As previously noted we can use the same ideas as above to similarly bound $\left|p_{\eta_{X,g'}}(t) - p_{\eta_{X,g}}(t)\right|$ for all $t \in \mathbb{R}$. It suffices to interchange X and Y in the above analysis. □

Lemma 4. *Let p_1, p_2 be two densities such that there exist $T, C > 1$ and $\alpha > 1$, for all $|t| > T$, $\max_{i\in[2]} p_i(t) < C\,|t|^{-\alpha}$. Suppose $\sup_{t\in\mathbb{R}} |p_1(t) - p_2(t)| < \epsilon$ for some $\epsilon < \min\left\{1/T^2, 1/(3e)\right\}$ satisfying the further condition: $\forall t > 1/\sqrt{\epsilon}$, $t^{(\alpha-1)/2} > \ln t$. We then have for ϵ sufficiently small*

$$|H(p_1) - H(p_2)| \leq 18\sqrt{\epsilon}\ln(1/3\epsilon) + \frac{4C\alpha}{\alpha - 1}\epsilon^{(\alpha-1)/4}.$$

Proof. For simplicity of notation in what follows, let $\tau \triangleq 1/\sqrt{\epsilon}$. Let $\mathcal{U} \triangleq [-\tau, \tau]$ and let $\mathcal{U}_{2>} \triangleq \{t \in \mathcal{U}, p_2(t) > 2\epsilon\}$. Define $\gamma(u) = -u\ln u$ for $u > 0$, and $\gamma(0) = 0$. We will use the fact that for the function $\gamma(\cdot)$ is increasing on $[0, 1/e]$. We have

$$\begin{aligned} H(p_1) &= \int_{\mathbb{R}\setminus\mathcal{U}} \gamma(p_1(t))\, dt + \int_{\mathcal{U}_{2>}} \gamma(p_1(t))\, dt \\ &\quad + \int_{\mathcal{U}\setminus\mathcal{U}_{2>}} \gamma(p_1(t))\, dt \\ &\leq \int_{\mathbb{R}\setminus\mathcal{U}} \gamma(p_1(t))\, dt + \int_{\mathcal{U}_{2>}} p_1(t)\ln\frac{1}{p_1(t)}\, dt \\ &\quad + \Sigma\left(\mathcal{U}\setminus\mathcal{U}_{2>}\right)\cdot\gamma(3\epsilon), \end{aligned} \tag{5}$$

since for $t \in \mathcal{U}\setminus\mathcal{U}_{2>}$ we have $p_1(t) \leq p_2(t) + \epsilon \leq 3\epsilon \leq 1/e$.

To bound the first term of (5), notice that

$$\begin{aligned} \int_\tau^\infty \gamma(p_1(t))\, dt &\leq \int_\tau^\infty -Ct^{-\alpha}\ln\left(Ct^{-\alpha}\right)\, dt \\ &\leq \int_\tau^\infty C\alpha t^{-\alpha}\ln t\, dt \\ &\leq \int_\tau^\infty C\alpha t^{-(\alpha+1)/2}\, dt \\ &\leq \frac{2C\alpha}{\alpha - 1}\tau^{-(\alpha-1)/2}, \end{aligned}$$

hence we have $\int_{\mathbb{R}\setminus\mathcal{U}} \gamma(p_1(t))\, dt \leq C'\tau^{-\alpha'}$, for $C', \alpha' > 0$.

Next we bound the second term of (5) as follows:

$$\begin{aligned}
&\int_{\mathcal{U}_{2>}} p_1(t) \ln \frac{1}{p_1(t)}\, dt \\
&\quad \leq \int_{\mathcal{U}_{2>}} (p_2(t)+\epsilon) \ln \frac{1}{p_2(t)-\epsilon}\, dt \\
&\quad = \int_{\mathcal{U}_{2>}} p_2(t) \ln \frac{1}{p_2(t)(1-\epsilon/p_2(t))}\, dt \\
&\quad + \int_{\mathcal{U}_{2>}} \epsilon \ln \frac{1}{p_2(t)-\epsilon}\, dt \\
&\quad \leq H(p_2) + \int_{\mathcal{U}_{2>}} p_2(t) \ln \frac{1}{1-\epsilon/p_2(t)}\, dt \\
&\quad + \int_{\mathcal{U}_{2>}} \epsilon \ln \frac{1}{\epsilon}\, dt \\
&\quad \leq H(p_2) + \int_{\mathcal{U}_{2>}} p_2(t) \ln(1+2\epsilon/p_2(t))\, dt \\
&\quad + \Sigma\left(\mathcal{U}_{2>}\right) \cdot \gamma(\epsilon) \\
&\quad \leq H(p_2) + 2\Sigma\left(\mathcal{U}_{2>}\right) \cdot \epsilon + \Sigma\left(\mathcal{U}_{2>}\right) \cdot \gamma(\epsilon).
\end{aligned}$$

Combining all the above, we have

$$\begin{aligned}
H(p_1) \leq & H(p_2) + 3\Sigma\left(\mathcal{U}\right) \cdot \gamma(3\epsilon) + C'\tau^{-\alpha'} \\
= & H(p_2) + 18\sqrt{\epsilon}\ln(1/3\epsilon) + C'\epsilon^{\alpha'/2}.
\end{aligned}$$

Notice that p_1 and p_2 are interchangeable in the above argument. The result therefore follows. □

We are now ready to prove the main theorem.

Proof of Theorem 1

Let $f(x) \triangleq \mathbb{E}\left[Y|x\right]$ and $g(y) \triangleq \mathbb{E}\left[X|y\right]$. By Lemma 1,

$$H(X) + H\left(\eta_{Y,f}\right) > H(Y) + H\left(\eta_{X,g}\right) + 8\epsilon, \quad (6)$$

for some $\epsilon > 0$.

Thus we detect the right direction $X \to Y$ if all quantities (a) $|H_n\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f}\right)|$, (b) $|H_n\left(\eta_{X,g_n}\right) - H\left(\eta_{Y,g}\right)|$, (c) $|H_n(X) - H(X)|$, and (d) $|H_n(Y) - H(Y)|$, are at most ϵ.

By assumption, (c) and (d) both tend to 0 in probability. The quantities (a) and (b) are handled as follows. We only show the argument for (a), as the argument for (b) is the same. We have:

$$\begin{aligned}
|H_n\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f}\right)| \leq & |H_n\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f_n}\right)| \\
& + |H\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f}\right)|.
\end{aligned}$$

Now $H_n\left(\eta_{Y,f_n}\right)$ is consistent for f_n fixed (it easy to check that $P_{\eta_{Y,f_n}}$ satisfies the necessary conditions provided f_n is bounded) and f_n is learned on an independent sample from H_n, we have $|H_n\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f_n}\right)| \xrightarrow{P} 0$.

By Lemma 3, convergence of f_n i.e. $\|f_n - f\|_{2,P_X} \xrightarrow{P} 0$ implies $\sup_t \left|p_{\eta_{Y,f_n}}(t) - p_{\eta_{Y,f}}(t)\right| \xrightarrow{P} 0$; this in turn implies by Lemma 4 that $|H\left(\eta_{Y,f_n}\right) - H\left(\eta_{Y,f}\right)| \xrightarrow{P} 0$.

Thus all quantities (a)-(d) are at most ϵ with probability going to 1. □

5.2. Coupled Regression and Residual-entropy Estimation

Here we consider a coupled version of the meta-algorithm where f_n and g_n are kernel regressors. This is described in the next subsection.

5.2.1. Kernel instantiation of the meta-algorithm

Regression: Although any kernel that is 0 outside a bounded region will work for the regression, we focus here (for simplicity) on the particular case where f_n and g_n are box-kernel regressors defined as follows (interchange X and Y to obtain $g_n(y)$):

$$f_n(x) = \frac{1}{n_{x,h}} \sum_{i=1}^{n} Y_i \mathbf{1}_{\{|X_i - x| < h\}}, \quad (7)$$

where $n_{x,h} = |i : |X_i - x| < h|$, for a bandwidth h.

Entropy estimation: Given a sequence $\epsilon = \{\epsilon_i\}_{i=1}^n$, and a bandwidth σ, define $p_{n,\epsilon}$ as follows:

$$\begin{aligned}
p_{n,\epsilon}(t) = & \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\epsilon_i - t}{\sigma}\right), \\
& \text{where } \int_{\mathbb{R}} K(u)\, du = 1, \left|\frac{d}{du}K(u)\right| < \infty, \\
& \text{and } K(u) = 0 \text{ for } |u| \geq 1.
\end{aligned}$$

Let $\epsilon_{Y,i} = Y_i - f_n(X_i)$ and $\epsilon_{X,i} = X_i - g_n(Y_i)$. The residual entropy estimators are defined as:

$$H_n\left(\eta_{Y,f_n}\right) \triangleq H\left(p_{n,\epsilon_Y}\right) \text{ and } H_n\left(\eta_{X,g_n}\right) \triangleq H\left(p_{n,\epsilon_X}\right). \quad (8)$$

5.2.2. Consistency result for coupled-estimation

We abuse notation and use h and σ to denote the bandwidth parameters used to estimate either f_n and $H_n\left(\eta_{Y,f_n}\right)$, or g_n and $H_n\left(\eta_{X,g_n}\right)$. We make the distinction clear whenever needed.

The consistency result depends on the following quantities bounded in Lemma 5.

Definition 6 (Expected average excess risk)**.** *Define* $R_n(f_n) \triangleq \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}|f_n(X_i) - f(X_i)|$ *and similarly* $R_n(g_n) \triangleq \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}|g_n(Y_i) - g(Y_i)|$.

We assume in this section that the noise η has exponentially decreasing tail:

Definition 7. *A r.v. Z has* **exponentially decreasing tail** *if there exists $C, C' > 0$ such that for all $t > 0$, $\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq Ce^{-C't}$.*

The following consistency theorem hinges on properly choosing the bandwidths parameters h and σ. Essentially we want to choose h such that regression estimation is consistent, and we want to choose σ so as not to overfit regression error. If the bandwidth σ is too small relative to regression error (captured by R_n), then the entropy estimator (for the residual entropy) is only fitting this error. The conditions on σ in the Theorem are mainly to ensure that σ is not too small relative to regression error R_n.

Theorem 2 (Coupled estimation)**.** *Suppose $X \xrightarrow{f,\eta} Y$ for some f, η, and suppose $P_{X,Y}$ satisfies Assumption 2, and η has exponentially decreasing tail. Let f_n, g_n, and H_n be defined as in Section 5.2.1, and let both $H_n(X)$ and $H_n(Y)$ be consistent as in Assumption 1.*

Suppose that :
*(*i*) For learning f_n and $H_n(\eta_{Y,f_n})$, we use $h = c_1 n^{-\alpha}$ for some $c_1 > 0$ and $0 < \alpha < 1$, and $\sigma = c_2 n^{-\beta}$ for some $c_2 > 0$ and $0 < \beta < \min\{(1-\alpha)/4, \alpha/2\}$.*
*(*ii*) For learning g_n and $H_n(\eta_{X,g_n})$, h satisfies $h \to 0$ and $nh \to \infty$, and σ satisfies $\sigma \to 0$, $n\sigma \to \infty$, and $\sigma = \Omega(R_n(g_n)^{-\gamma})$ for some $0 < \gamma < 1/2$.*

Then the probability of correctly detecting $X \to Y$ goes to 1 as $n \to \infty$.

The theorem relies on Lemma 5 which bounds the errors R_n for both f_n and g_n. Suppose $X \xrightarrow{f,\eta} Y$, then if f is smooth or continuously differentiable, $R_n(f_n) \to 0$, and in fact we can obtain finite rates of convergence for $R_n(f_n)$, thus yielding advice on setting σ. The second part of the Lemma corresponds to this situation.

However, as mentioned earlier in the paper introduction, a smooth f does not ensure that $g(y) \triangleq g(X|y)$ is smooth or even continuous, so we do not have rates for $R_n(g_n)$. We can nonetheless show that $R_n(g_n)$ would generally converge to 0, which is sufficient for there to be proper settings for σ (i.e. σ larger than the error, but also tending to 0).

We note that the r.v.'s X and Y are interchangeable in this lemma since it does not assume $X \to Y$. The proof is given in the extended version of the paper.

Lemma 5. *Let f_n be defined as in (7). Let $f(x) \triangleq \mathbb{E}[Y|x]$. Suppose (*i*) $\mathbb{E}Y^2 < \infty$ and that f is bounded; $h \to 0$ and $nh \to \infty$. Then $\mathbb{E}\frac{1}{n}\sum_1^n |f_n(X_i) - f(X_i)| \xrightarrow{n\to 0} 0$.*

*Suppose further (*ii*) that P_X has bounded support and that f is continuously differentiable; $h = c_1 n^{-\alpha}$ for some $c_1 > 0$ and $0 < \alpha < 1$. Then we have $\mathbb{E}\frac{1}{n}\sum_1^n |f_n(X_i) - f(X_i)| \leq c_2 n^{-\beta}$, for $\beta \triangleq \min\{(1-\alpha)/2, \alpha\}$.*

The main theorem of this sectin is proved in the long version of this paper.

6. Final Remarks

We derived the first consistency results for an existing family of procedures for causal inference under the Additive Noise Model. We obtained mild algorithmic requirements, and various distributional tail conditions which guarantee consistency. The present work focuses on the case of two r.v.s X and Y, which captures the inherent difficulties of consistency. We believe however that the insights developed should extend to the case of random vectors under corresponding tail conditions. The details however are left for future work.

Another interesting multivariate situation is that of a causal network of r.v.s. as in Peters et al. (2011b) dicussed earlier. Extending our consistency results to this particular multivariate case would primarily consist of extending our distributional tail conditions to the tails of distributions resulting from conditioning on appropriate sets of variables in the network. This is however a non-trivial extension as it involves, e.g. for the convergence of conditional entropies, some additional integration steps that have to be carefully worked out.

A possible future direction of investigation is to understand under what conditions finite sample rates can be obtained for such procedures. For reasons explained earlier, we do not believe that this is possible without less general distributional assumptions.

References

Beirlant, Jan, Dudewicz, Edward J, Györfi, László, and Van der Meulen, Edward C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–40, 1997.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *arXiv:1310.1533*, 2013.

Cover, Thomas M, Thomas, Joy A, and Kieffer, John. Elements of information theory. *SIAM Review*, 36(3):509–510, 1994.

Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution Free Theory of Nonparametric Regression*. Springer, New York, NY, 2002.

Hoyer, Patrik O, Janzing, Dominik, Mooij, JM, Peters, Jonas, and Schölkopf, Bernhard. Nonlinear causal dis-

covery with additive noise models. *Proceedings of Advances in Neural Processing Information Systems*, 2009.

Hyvärinen, Aapo, Shimizu, Shohei, and Hoyer, Patrik O. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the 25th international conference on Machine learning*, pp. 424–431. ACM, 2008.

Nowzohour, Christopher and Bühlmann, Peter. Score-based causal learning in additive noise models. *arXiv preprint arXiv:1311.6359*, 2013.

Pearl, Judea. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

Peters, Jonas, Janzing, Dominik, and Scholkopf, Bernhard. Causal inference on discrete data using additive noise models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2436–2450, 2011a.

Peters, Jonas, Mooij, Joris, Janzing, Dominik, and Schölkopf, Bernhard. Identifiability of causal graphs using functional models. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011b.

Shimizu, Shohei, Hoyer, Patrik O, Hyvärinen, Aapo, and Kerminen, Antti. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Spirtes, Peter, Glymour, Clark N, and Scheines, Richard. *Causation Prediction & Search 2e*, volume 81. MIT press, 2000.

Tillman, Robert, Gretton, Arthur, and Spirtes, Peter. Nonlinear directed acyclic structure learning with weakly additive noise models. *Proceedings of Advances in Neural Processing Information Systems*, 22:1847–1855, 2009.

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.