Journal of Machine Learning Research 16 (2015) 1923-1948 Submitted 6/15; Published 9/15

Semi-Supervised Interpolation in an Anticausal Learning Scenario

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems
Spemannstr. 38
72076 Tübingen, Germany

Editor: Alex Gammerman and Vladimir Vovk

Abstract

According to a recently stated 'independence postulate', the distribution P_{cause} contains no information about the conditional $P_{\text{effect}|\text{cause}}$ while P_{effect} may contain information about $P_{\text{cause}|\text{effect}}$. Since semi-supervised learning (SSL) attempts to exploit information from P_X to assist in predicting Y from X, it should only work in anticausal direction, i.e., when Y is the cause and X is the effect. In causal direction, when X is the cause and Y the effect, unlabelled x-values should be useless. To shed light on this asymmetry, we study a deterministic causal relation $Y = f(X)$ as recently assayed in Information-Geometric Causal Inference (IGCI). Within this model, we discuss two options to formalize the independence of P_X and f as an orthogonality of vectors in appropriate inner product spaces. We prove that unlabelled data help for the problem of interpolating a monotonically increasing function if and only if the orthogonality conditions are violated – which we only expect for the anticausal direction. Here, performance of SSL and its supervised baseline analogue is measured in terms of two different loss functions: first, the mean squared error and second the surprise in a Bayesian prediction scenario.

Keywords: semi-supervised learning, anticausal learning, independence of cause and mechanism, information geometry, causality

1. Introduction

Semi-supervised learning (SSL) has received increasing attention during the past decade (Darnstädt et al., 2013; Ben-David et al., 2008; Yuanyuan et al., 2010; Chapelle et al., 2006). In contrast to supervised learning, where the prediction of a variable Y from another variable X is based on pairs $(x_1, y_1), \ldots, (x_n, y_n)$, semi-supervised learning uses additional x-values $x_{n+1}, \ldots, x_{n+m}$ to improve the prediction. Motivated by the fact that the y-values are often discrete variables, that is, 'labels', one often talks about the pairs as *labelled* instances and the unpaired x-values as *unlabelled* ones.

One can easily imagine scenarios where labelled instances are rare and unlabelled ones are easily available: consider, for example, the task of text classification, where labelling has to be done by humans while unlabelled instances can be retrieved from the internet automatically. Hence, SSL is useful provided that the unlabelled x-values indeed contain information about the relation between X and Y. Given the standard scenario where the pairs are i.i.d. drawn from P_{XY} and the unlabelled x-values from the corresponding marginal distribution P_X, the essential question is the following. Predicting Y from X amounts to knowing properties of $P_{Y|X}$, while the unlabelled x-values only tell us something about P_X. Why should P_X contain information about $P_{Y|X}$?

Some recent approaches to distinguish cause and effect in causal structure learning (Janzing and Schölkopf, 2010; Daniusis et al., 2010; Janzing et al., 2012; Sgouritsa et al., 2015) were motivated by an informal 'independence' postulate stating that P_{cause} and $P_{\text{effect}|\text{cause}}$ contain no information about each other. On the other hand, P_{effect} and $P_{\text{cause}|\text{effect}}$ may contain information about each other. This has been shown by means of several toy examples (Janzing and Schölkopf, 2010; Daniusis et al., 2010; Janzing et al., 2012) using appropriate formalizations of the independence postulate. In the same spirit, Schölkopf et al. (2012, 2013) argue that under the independence postulate, SSL cannot work in the causal setting, that is, if X is the cause and Y the effect (provided that there is no common cause of both), while it may work in anticausal setting, i.e., when the cause is predicted from the effect. In a typical scenario of SSL that often appears in the literature (Chapelle et al., 2006), Y attains few values $\{1, \ldots, k\}$ only (Zhang and Oles, 2000) and $X \in \mathbb{R}^d$ is a high-dimensional vector. Then different labels j may correspond to different clusters in $\mathbb{R}^d$. If they are sufficiently apart, the modes of P_X tell us the centers of the clusters, which helps in learning $P_{Y|X}$ from fewer data. Distributions that satisfy this (loose) condition are said to follow the cluster assumption, a case for which SSL can plausibly be justified (Chapelle et al., 2006): as long as each cluster contains some labelled data points, we can propagate the labels to the other points in the same cluster, and thus convert the semi-supervised learning problem to a supervised one. In our terminology, this assumption implies that points in the same cluster have the same label, i.e., certain properties of P_X imply properties of $P_{Y|X}$. A related assumption states that the separating boundary should lie in a region of low density of P_X (Chapelle et al., 2006) – again, an assumption relating P_X and $P_{Y|X}$.

The goal of this paper is to provide a mathematical understanding of why the performance of SSL is related to the causal direction. Previous work remains vague regarding the question in what sense P_{effect} may contain information about $P_{\text{cause}|\text{effect}}$ and which mathematical postulates about asymmetries between cause and effect are needed for this claim. Here we present a model in which a well-defined independence assumption between P_{cause} and $P_{\text{effect}|\text{cause}}$ ensures that unlabelled data from the effect help in the sense of quantitatively improving the prediction of the cause from the effect with respect to a natural loss function, while it does not help in causal direction. To this end, we have chosen a model where X and Y have the same range. The more popular case where X is high-dimensional and Y of lower dimension or even a discrete label could be misleading for our purposes: different ranges define an asymmetry between X and Y that could erroneously be attributed to the fact that one is the cause and the other the effect.

We study the following simple **interpolation problem**: *Let X and Y be random variables attaining values in $[0,1]$, deterministically related by $Y = f(X)$, where f is an unknown bijective strictly monotonically increasing map. We are given $n-1$ points $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. For some additional x-value x_n, we seek to infer the corresponding y-value $y_n = f(x_n)$.*

We will analyze why knowing P_X enables a better estimation (which implies that P_X and $P_{Y|X}$ are somehow dependent), given that a certain independence between P_Y and $P_{X|Y}$ holds.

The paper is structured as follows. Section 2 introduces a toy model of a bijective deterministic relation between X and Y and formalizes independence between P_{cause} and $P_{\text{effect}|\text{cause}}$ in two different ways. We explain why this independence implies dependence between P_{effect} and $P_{\text{cause}|\text{effect}}$ with respect to both formalizations. Section 3 describes the interpolation problem in the supervised scenario (i.e., with no unlabelled points) and presents a straightforward solution via linear interpolation, which will be the baseline our SSL method is later compared to.

Section 4 describes a semi-supervised modification and shows that the advantage can be quantified in terms of the dependence measures introduced in Section 2. The main contribution of this paper is to describe the relation between the performance of SSL to a mathematically well-defined notion of dependence between P_X and $P_{Y|X}$. Although our toy scenario is certainly an oversimplification compared to real SSL scenarios, the value of this work lies in providing the first link between causal direction and applicability of SSL that can be proven, subject to an assumption that links causality to statistics.

2. Asymmetries Between Cause and Effect for Deterministic Relations

Our restriction to monotonically increasing bijections of $[0,1]$ coincides with the typical toy scenario used by Daniusis et al. (2010); Janzing et al. (2012) to explain Information-Geometric Causal Inference (IGCI) although the formalism of IGCI introduced therein is actually more general.

We are given two random variables C, E ('cause' and 'effect') attaining values in $[0,1]$. We assume that their distributions P_C and P_E have strictly positive densities p_C and p_E with respect to Lebesgue measure. We will often use $p(c)$ as short hand for $p_C(c)$, for instance. Assume we observe that C and E are deterministically related by

$$E = g(C) \quad \text{and} \quad C = g^{-1}(E)\,,$$

for some strictly monotonically increasing diffeomorphism[1] g of $[0,1]$.

So far, the assumptions are symmetric with respect to C and E and there is no reason why observing the joint distribution of E and C should enable one to infer which variable is the cause and which the effect, assuming that exactly one of the alternatives is true. The problem of distinguishing cause and effect gets solvable only after introducing an assumption that links the causal direction to an observable implication. The essential idea is that g (which uniquely determines $P_{E|C}$) and p_C do not contain information about each other. Subsections 2.1 and 2.2 will describe two different formalizations of this idea which are the basis for two different SSL methods presented in Subsections 4.1 and 4.2, respectively.

2.1 Uncorrelatedness Between p_C and Slope

To formalize the idea of independence between g and p_C, Daniusis et al. (2010); Janzing et al. (2012, 2015) postulate uncorrelatedness between p_C and the logarithm of the derivative of g, which will be explained in Subsection 2.2. Here we state an assumption that simplifies the former by dropping the logarithm:

[1] The 'diffeomorphism' assumption is convenient for the theory although it can be significantly weakened. The example in Figure 1(a) uses functions g and g^{-1} that are almost everywhere differentiable, which is also sufficient.

Independence Assumption 1 (with slope) *If C causes E with $E = g(C)$ then*

$$\mathrm{Cov}[g', p_C] = 0\,. \tag{1}$$

Here, both functions g' and p_C are considered as random variables on the probability space $[0,1]$ with Lebesgue measure. Their covariance, i.e., the left hand side of (1), equals

$$\int_0^1 g'(c)p(c)dc - \int_0^1 g'(c)dc \int_0^1 p(c)dc = \int_0^1 g'(c)p(c)dc - 1\,. \tag{2}$$

It turns out that Independence Assumption 1 implies that P_E contains information about g^{-1} (and thus about $P_{C|E}$):

Lemma 1 (p_E correlates with slope) *Let $g \neq id$ and (1) hold. Then the derivative of g^{-1}, denoted by $g^{-1'}$, is positively correlated with p_E:*

$$\mathrm{Cov}[g^{-1'}, p_E] > 0\,. \tag{3}$$

Proof By substitution of variables, (2) implies

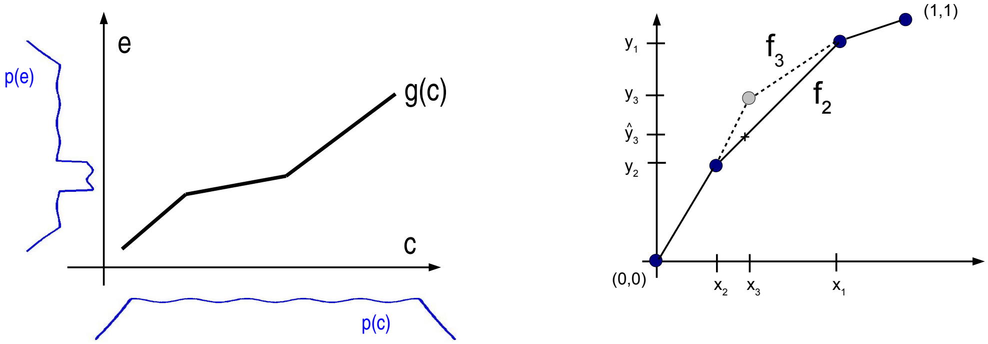
$$\int_0^1 p(e)\frac{1}{g^{-1'}(e)}de = 1\,. \tag{4}$$

We then conclude

$$\begin{aligned}\int_0^1 p(e)g^{-1'}(e)de &= \int_0^1 p(e)g^{-1'}(e)de \cdot \int_0^1 p(e)\frac{1}{g^{-1'}(e)}de \\ &= \int_0^1 p(e)\left(\sqrt{g^{-1'}(e)}\right)^2 de \cdot \int_0^1 p(e)\left(\frac{1}{\sqrt{g^{-1'}(e)}}\right)^2 de \\ &\geq \left(\int_0^1 p(e)\sqrt{g^{-1'}(e)}\frac{1}{\sqrt{g^{-1'}(e)}}de\right)^2 = 1\,,\end{aligned}$$

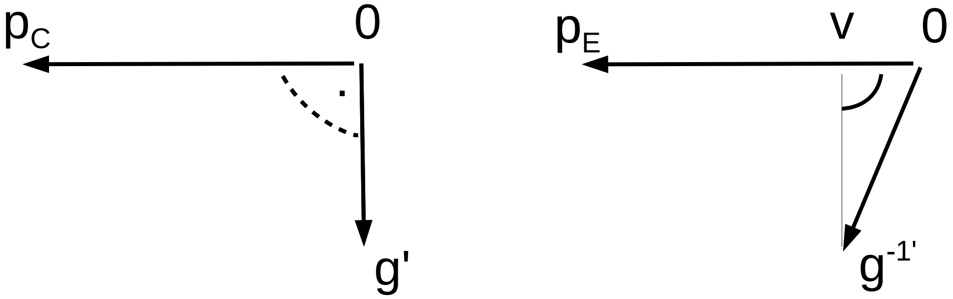
where we have applied the Cauchy-Schwarz inequality to the inner product $\langle \cdot, \cdot \rangle = \int p(e) \cdot \cdot de$ (note that it is strictly positive because p_E is strictly positive). Therefore we only have equality if $\sqrt{g^{-1'}}$ and $1/\sqrt{g^{-1'}}$ are linearly dependent, i.e., g' is constant and thus g is the identity due to $g(0) = 0$ and $g(1) = 1$. ■

Figure 1(a) provides a first intuition about Lemma 1: whenever the slope of g has been chosen independently of p_E, the density p_C tends to be high in regions where g is flat and g^{-1} is steep. Figures. 2(a) and 2(b) visualize the geometric content of Lemma 1 in the following sense. The covariance defines an inner product in the space of square integrable random variables if variables are identified up to constants. Then we have postulated orthogonality of g' and p_C and concluded non-orthogonality of g^{-1} and p_E. Therefore, the projection v of g^{-1} onto the line $(0, p_E)$ is closer to g^{-1} than 0. Within our setting, this point v will later play a crucial role for constructing the optimal prediction of g^{-1} that can be obtained from p_E.



(a) (b)

Figure 1: (a) If g has been designed independently of p_C, then the density p_E tends to be high in regions where g is flat. Source: Janzing et al. (2012). (b) The piecewise linear function f_2 interpolating the observations $(x_1, y_1), (x_2, y_2)$ is used for predicting y_3. f_3 accounts also for the point (x_3, y_3) and is later used to predict y_4 once x_4 is provided.



(a) (b)

Figure 2: Orthogonality of the random variables p_C and g' (in the sense of vanishing covariance) in Figure (a) implies non-orthogonality of p_E and $g^{-1'}$ in Figure (b). In Subsection 4.1, the squared distance of v and 0 will be the amount by which SSL can improve the performance of the interpolation.

2.2 Uncorrelatedness Between p_C and Logarithmic Slope

To phrase independence of g and p_C as uncorrelatedness of p_C and the derivative of g is certainly only one simple choice out of many options. Instead, Daniusis et al. (2010); Janzing et al. (2012) postulate uncorrelatedness between p_C and the *logarithm* of the derivative of g:

Independence Assumption 2 (with logarithmic slope) *If C causes E with $E = g(C)$ then*

$$\mathrm{Cov}[\log g', p_C] = 0\,. \tag{5}$$

Here, both functions $\log g'$ and p_C are considered as random variables on the probability space $[0,1]$. Again, their covariance is then computed with respect to the Lebesgue measure, i.e., the left hand side of (5) is short hand for

$$\int_0^1 \log g'(c)p(c)dc - \int_0^1 \log g'(c)dc \int_0^1 p(c)dc = \int_0^1 \log g'(c)p(c)dc - \int_0^1 \log g'(c)dc\,.$$

Assumption 2 admits several information theoretic interpretations (Daniusis et al., 2010; Janzing et al., 2012, 2015) of which we only explain the ones that are required for our analysis.

It turns out (Daniusis et al., 2010, Section 2) that Assumption 2 implies that P_E contains information about g^{-1} (and thus about $P_{C|E}$):

Lemma 2 (p_E correlates with logarithmic slope) *Let $g \neq id$ and (5) hold. Then the logarithm of the derivative of g^{-1}, denoted by $g^{-1'}$, is positively correlated with p_E:*

$$\mathrm{Cov}[\log g^{-1'}, p_E] > 0\,. \tag{6}$$

Our algorithm and the performance analysis will be based on the following information geometric rephrasing of the above.

Lemma 3 (covariance as difference of relative entropies) *Let*

$$D(q\|r) := \int_0^1 q(w) \log \frac{q(w)}{r(w)} dw$$

denote the relative entropy distance between the probability densities q and r. Then,

$$\mathrm{Cov}[\log g', p_C] = -D(p_C\|g') + D(p_C\|u) + D(u\|g')\,,$$

where u denotes the uniform density. Here we have interpreted g' as probability density which is possible due to $g' > 0$ and $\int g'(c)dc = 1$.

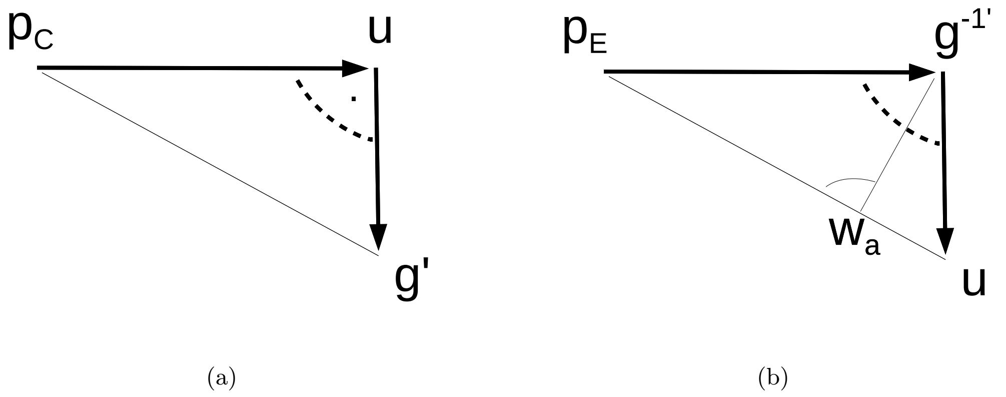
The following conclusion is immediate:

Corollary 1 (independence as orthogonality in information space) *(5) is equivalent to*

$$D(p_C\|g') = D(p_C\|u) + D(u\|g')\,. \tag{7}$$

Likewise, (6) is equivalent to

$$D(p_E\|g^{-1'}) < D(p_E\|u) + D(u\|g^{-1'})\,. \tag{8}$$



(a) (b)

Figure 3: (a) Independence Assumption 2 for P_C and $P_{E|C}$ implies that (p_C, u, g') is a Pythagorean triple, i.e., there is a rectangle at u. (b) Since bijections preserve relative entropy, the right angle for the backward direction occurs at $g^{-1'}$ instead of u, as would be required by the corresponding independence assumption for P_E and $P_{C|E}$. The point w_a obtained by projecting $g^{-1'}$ onto the line u, p_E will later play a crucial role for our SSL method and the distance $D(w_a\|u)$ will quantify the amount by which SSL improves the interpolation.

Without going to the details of information geometry Amari and Nagaoka (1993), we use some of its terminology and mention that due to (7), (p_C, u, g') is called a Pythagorean triple. This is visualized by drawing a right angle at u, see Figure 3(a). The idea is that square distance in Euclidean geometry is replaced with relative entropy in information geometry and therefore (7) replaces the usual Pythagorean theorem.[2] This way, Assumptions 1 and 2 both amount to orthogonality conditions in appropriate spaces.

Since relative entropy is preserved under bijections, we also have:

Lemma 4 (right angle at $g^{-1'}$) *Eq. (5) is equivalent to*

$$D(p_E\|u) = D(p_E\|g^{-1'}) + D(g^{-1'}\|u)\,. \tag{9}$$

Geometrically, this means that the right angle now occurs at $g^{-1'}$, as visualized by Figure 3(b), whereas independence between p_E and $g^{-1'}$ would require it to occur at u. In other words, by formalizing independence between input distribution and function as a certain orthogonality in information space, independence in causal direction implies dependence in anticausal direction. IGCI uses this asymmetry for inferring which of the two variables is the cause.

The goal of this paper is to answer the question why P_X is helpful for the interpolation problem stated in Section 1 when $X = E$ and $Y = C$, while it is useless when $X = C$

[2] Then, the m-geodesic connecting p_C and u (given by the line $\lambda p_C + (1-\lambda)u$) is orthogonal to the e-geodesic connecting u and g' which is given by an affine combination on the logarithmic scale, that is, by $\lambda \log u + (1-\lambda)\log g'$.

and $Y = E$. Some thoughts on this can be found in Janzing et al. (2015, Section 4), but here we will describe a learning scenario where the information of P_X on f amounts to reducing the loss with respect to some natural loss function. To this end, we first describe a baseline method for the interpolation problem in Section 3 and analyze its performance with respect to two different loss functions. In Section 4 it will turn out that our two different formalizations of dependence vs. independence in Subsections 2.1 and 2.2 yield two different algorithms each of which improves the performance with respect to one of these loss functions.

3. Baseline Solutions of the Interpolation Problem

To analyze the performance of our interpolation methods (baseline and SSL) we consider a game consisting of infinitely many steps: In the nth step, we are given $(n-1)$ pairs

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$$

obtained by i.i.d. sampling from P_{XY}. After observing the next x-value x_n, we are supposed to infer the corresponding value y_n. Having inferred it, we are told the true value y_n and the next x-value x_{n+1}. The reason why we define this game is that our theory will not provide a performance statement for any *specific* n. Instead, we will show that SSL outperforms the baseline method on average over all n until some $n_{\max}$ if $n_{\max}$ tends to infinity. Note, however, that the first step $n = 1$ would be usually called ‘unsupervised learning’, which we include as special case of SSL in our analysis.

Note, moreover, that ‘inferring y_n’ can mean two different things: either one infers one specific value $\hat{y}_n$. Then the performance is evaluated by some distance measure between the estimated value $\hat{y}_n$ and the true value y_n. The other sense of ‘inferring’ is to define some conditional probability density[3]

$$\mathrm{pr}(y_n | x_1, \dots, x_n, y_1, \dots, y_{n-1}) \tag{10}$$

expressing one’s belief about y_n. Then it is natural to evaluate the performance of the prediction by the ‘surprise’ given by the negative logarithm of (10). Subsections 3.1 and 3.2 describe the supervised baseline scenarios for the two different settings.

3.1 Predicting One Specific Value by Linear Interpolation

As baseline method we consider interpolation by piecewise linear functions:

Definition 1 (linear interpolation) *For some $(n-1)$-tuple of points*

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1}), \quad \textit{with } n \geq 1,$$

let f_n denote the function that linearly interpolates between these points (see Figure 1(a), right). Explicitly, it is given by first ordering the x-values $x_1^\circ < \dots < x_{n-1}^\circ$, which also

[3] We use the notation pr to indicate that it is not connected to the probability densities p_X and p_Y. In a fully Bayesian scenario we would parameterize the set of distributions P_{cause} and the set of functions g and then define a prior on both parameter spaces. Here, pr expresses a belief on y_n that will later be based on some naive smoothness assumption formalized by the Dirichlet prior without accounting for any explicit generating model.

orders the y-values $y'_1 < \cdots < y'_{n-1}$. Then f_{n-1} is the piecewise linear function that linearly connects (x°_j, y°_j) with $(x^\circ_{j+1}, y^\circ_{j+1})$ for $j = 0, \ldots, n-1$ after we have set $x^\circ_0 = y^\circ_0 = 0$ and $x^\circ_n = y^\circ_n = 1$. Hence, f_0 is the identity.

Although the interpolating function f_n depends on the whole $(n-1)$-tuple of points, it will be convenient to have only the index n since we refer to a fixed list of observations (obtained by i.i.d sampling from P_{XY}) of which we only know the first $n-1$. We set

$$\hat{y}_n := f_{n-1}(x_n)\,,$$

with f_{n-1} as in Definition 1, see also Figure 1(a), right. Here and throughout the paper i will denote the index for which x_n lies in the interval $(x^\circ_i, x^\circ_{i+1})$ (see the notation of Definition 1). Then the estimated value is explicitly given by

$$\hat{y}_n = \frac{x_n - x^\circ_i}{x^\circ_{i+1} - x^\circ_i}(y^\circ_{i+1} - y^\circ_i) + y^\circ_i\,. \tag{11}$$

To analyze the performance of the SSL version versus standard liner interpolation, it would be natural to measure the deviation of $\hat{y}_n$ from y_n via the usual squared loss $(\hat{y}_n - y_n)^2$. Here we modify this term as follows:

Definition 2 (modified squared loss) *The deviation between the estimated value $\hat{y}_n$ and the true value y_n in step n is measured by the loss*

$$L_n(y_n, \hat{y}_n) := \left(\frac{1}{x_n - x^\circ_i} + \frac{1}{x^\circ_{i+1} - x_n}\right)(\hat{y}_n - y_n)^2\,, \tag{12}$$

where i again denotes the index for which $x_n \in (x^\circ_i, x^\circ_{i+1})$.

The additional weighting factor amounts to stronger penalizing the deviation for those cases where x_n is close to the neighbors x°_i and x°_{i+1}. This can be justified by the idea that these errors should count stronger because one should actually be able to infer y_n more accurately when labelled points are close. The main reason, however, for the weighting factor is that it is necessary to link the performance of linear interpolation to Independence Assumption 1. The following reinterpretation will later be the reason why the loss (12) is convenient for our purposes:

Lemma 5 (squared loss as distance of derivatives) *Let $\hat{f}_n$ and f_n be the piecewise linear functions (linear on our n intervals) that interpolate the points $(x_n, \hat{y}_n)$ and (x_n, y_n), respectively, in addition to the points (x_i, y_i) for $i = 1, \ldots, n-1$. Then,*

$$L_n(y_n, \hat{y}_n) = \int_0^1 (f'_n(x) - \hat{f}'_n(x))^2 dx\,. \tag{13}$$

Proof:

$$\int_0^1 (f'_n(x) - \hat{f}'_n(x))^2 dx = \left(\frac{y_n - y^\circ_i}{x_n - x^\circ_i} - \frac{\hat{y}_n - y^\circ_i}{x_n - x^\circ_i}\right)^2 (x_n - x^\circ_i) +$$

$$\left(\frac{y^\circ_{i+1} - y_n}{x^\circ_{i+1} - x_n} - \frac{y^\circ_{i+1} - \hat{y}_n}{x^\circ_{i+1} - x_n}\right)^2 (x^\circ_{i+1} - x_n) = (y_n - \hat{y}_n)^2 \left(\frac{1}{x_n - x^\circ_i} + \frac{1}{x^\circ_{i+1} - x_n}\right) = L_n(y_n, \hat{y}_n)\,.$$

■

We now show that the loss until step $n_{\max}$ and the total loss over infinitely many steps can be given in a concise form. The proofs will be skipped because the corresponding results for the SSL scenario (Lemma 16 and Theorem 2) contain the statements below as special cases.

Lemma 6 (total loss until step $n_{\max}$) *The sum over all modified quadratic errors reads:*

$$\sum_{n=1}^{n_{\max}} L_n(y_n, \hat{y}_n) = \int_0^1 (f'_{n_{\max}}(x) - 1)^2 dx\,.$$

Therefore, the asymptotic loss reads:

Lemma 7 (total loss) *The sum over all modified quadratic errors reads:*

$$\sum_{n=1}^{\infty} L_n(y_n, \hat{y}_n) = \int_0^1 (f'(x) - 1)^2 dx = \mathrm{Var}(f')\,,$$

where we consider f' as random variable on the probability space $[0,1]$ with respect to the Lebesgue measure.

Recall that we have already considered derivatives of functions as random variables in Subsection 2.1. It is intuitively plausible that the complexity of the interpolation problem depends on the non-linearity of f, which can be quantified by the variance of f'. Note that this variance is also the squared length of the vectors g' and $g^{-1'}$ in Figure 2(a). Hence, we have linked the modified quadratic errors to Euclidean geometry in the space of random variables of Subsection 2.1. Accordingly, the non-orthogonality of p_E and $g^{-1'}$ in this space will be employed to construct an SSL algorithm that outperforms linear interpolation with respect to the modified quadratic errors.

3.2 Interpolation via a Dirichlet Process

To obtain a probability distribution that expresses our belief about y_n, given $x_1, \ldots, x_n$ and $y_1, \ldots, y_{n-1}$, we define a prior over the monotonically increasing functions. An arbitrary monotonic function f on $[0,1]$ with $f(0) = 0$ and $f(1) = 1$ can be interpreted as cumulative distribution function of a probability distribution on $[0,1]$. Since Dirichlet distributions can be used as priors for probability distributions, it is therefore also natural to use them as priors for increasing functions. We first introduce Dirichlet distributions of finite order (Balakrishnan and V., 2003):

Definition 3 (Dirichlet distribution) *The Dirichlet distribution* $\mathrm{Dir}(\alpha)$ *of order k and parameter vector $\alpha = (\alpha_1, \ldots, \alpha_k)$ with $\alpha_j > 0$ is defined as the density on the simplex*

$$\left\{ \theta \in \mathbb{R}^k \,\middle|\, \theta_j > 0, \sum_{j=1}^{k} \theta_j = 1 \right\}\,,$$

given by

$$\mathrm{pr}(\theta) := \frac{1}{B(\alpha)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}\,, \tag{14}$$

where $B(\alpha)$ is the normalization constant

$$B(\alpha) := \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j)}\,,$$

and Γ denotes the gamma function.

The following known result shows how the α_j control the expectations:

Lemma 8 (expectation of Dirichlet distribution) *The expectation of each θ_j is given by*

$$\mathbf{E}[\theta_j] = \frac{\alpha_j}{\sum_{j=1}^{k} \alpha_j}\,.$$

The sum over all α_j then controls to what extent the distribution is concentrated around its mean. The following well-known property will be crucial below:

Lemma 9 (aggregation property of Dirichlet) *If $(\theta_1, \ldots, \theta_k)$ is a random vector distributed according to $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ then $(\theta_1, \ldots, \theta_{k-2}, \theta_{k-1} + \theta_k)$ is distributed according to*
$\mathrm{Dir}(\alpha_1, \ldots, \alpha_{k-2}, \alpha_{k-1} + \alpha_k)$.

When a Dirichlet distribution is used to describe a distribution over distributions, then θ is the probability vector of k events. If we define Δ_j^Y with $j = 0, \ldots, n$ as the gaps obtained by ordering all values $y_1, \ldots, y_n$, then $\sum_{j=0}^{n} \Delta_j^Y = 1$ and thus the Dirichlet distribution of order $n+1$ defines a distribution over the set of possible difference vectors $\Delta^Y := (\Delta_0^Y, \ldots, \Delta_n^Y)$. However, we have to define distributions of order n for arbitrary n and need to ensure that the distributions defined for different n are consistent in the sense that marginalizing the distribution of $y_1, \ldots, y_n$ over y_n coincides with the distribution of $y_1, \ldots, y_{n-1}$ we define for $\tilde{n} = n - 1$. To this end, we use a Dirichlet process, which is the generalization of a Dirichlet distribution to infinite order:

Definition 4 (prediction via Dirichlet process) *Given the values $x_1, \ldots, x_n$, we define the probability density for the corresponding y-values by*

$$\mathrm{pr}(y_1, \ldots, y_n | x_1, \ldots, x_n) = \frac{1}{B(\alpha)} \prod_{j=0}^{n} (\Delta_j^Y)^{\alpha_j - 1}\,, \tag{15}$$

where the parameters are defined via the gaps of the corresponding x-values:

$$\alpha_j := \lambda \Delta_j^X \quad j = 0, \ldots, n\,, \tag{16}$$

where Δ_j^X are defined in analogy to Δ_j^Y. Here, $\lambda > 0$ is a parameter that controls to what extent we prefer linear function[4]

[4] It should be noted that functions obtained by a Dirichlet process are almost surely discontinuous (Blackwell, 1973) although we have assumed the true function f to be differentiable. Yet, the process defines a reasonable prior for our 'naive' prediction scheme of finitely many y-values. Later, we will let λ go to infinity (which renders the discontinuities arbitrarily small) *before* we consider $n \to \infty$.

To understand this Definition, we need a few remarks. First note that actually Δ^Y is Dirichlet distributed, but the same probability density can be used for $\mathbf{y} := (y_1, \dots, y_n)$ since the Jacobian of the transformation from Δ^Y to $\mathbf{y}$ is 1. This shows that the normalization of (14) still remains correct. To choose the parameters α_j proportional to the gaps in x-direction (16) amounts to taking the uniform distribution as 'base measure' according to standard terminology of Dirichlet processes. We will later see that changing the base measure provides a simple way to define an SSL version of the above prediction. Lemma 8 shows the implication of this choice: the expectation of each Δ_j^Y is given by the corresponding gap Δ_j^X. In this sense, the Dirichlet process a priori favors the linear function. For our further analysis it is also important to note that Lemma 9 implies

$$\mathrm{pr}(y_1, \dots, y_{n-1} | x_1, \dots, x_n) = \mathrm{pr}(y_1, \dots, y_{n-1} | x_1, \dots, x_{n-1}) . \tag{17}$$

Hence, using (14) for n points and marginalizing over y_n is the same as applying it to $\tilde{n} := n - 1$ points only, which is the sense of consistency we have demanded above. In other words, the unlabelled value x_n is irrelevant for the prediction of the remaining $(n-1)$ y-values.

After having seen $(n-1)$ points, we interpolate via the prediction rule

$$\mathrm{pr}(y_n | x_1, \dots, x_n, y_1, \dots, y_{n-1}) = \frac{\mathrm{pr}(y_1, \dots, y_{n-1}, y_n | x_1, \dots, x_n)}{\mathrm{pr}(y_1, \dots, y_{n-1} | x_1, \dots, x_n)} . \tag{18}$$

Although our performance analysis does not require the explicit form of the left hand side of (18), the following result (which is shown in Appendix A) provides a better understanding about what it does:

Lemma 10 (interpolation by Dirichlet of order 2**)** *Eq. (15) yields*

$$\mathrm{pr}(y_n | x_1, \dots, x_n, y_1, \dots, y_{n-1}) = \frac{1}{(y_{i+1}^\circ - y_i^\circ) B(\alpha)} \prod_{l=1}^{2} (\theta_l)^{\alpha_l - 1} ,$$

with $\theta_1 := (y_n - y_i^\circ)/(y_{i+1}^\circ - y_i^\circ)$ *and* $\theta_2 := 1 - \theta_1$. *The parameter vector reads*

$$\alpha := \lambda((x_n - x_i^\circ), (x_{i+1}^\circ - x_n)) .$$

Note that we need the additional normalization factor $(y_{i+1}^\circ - y_i^\circ)$ compared to (14) because the Dirichlet distribution is actually a normalized probability density for $\theta_1 \in (0, 1)$ which we have transformed into a density for $y_n \in (y_{i+1}^\circ, y_i^\circ)$. Due to Lemma 8 the expectation of the ratio $\theta_1 = (y_n - y_i^\circ)/(y_{i+1}^\circ - y_i^\circ)$ is thus given by the corresponding ratio $(x_n - x_i^\circ)/(x_{i+1}^\circ - x_i^\circ)$. Hence, (18) favors piecewise linear interpolation as defined in Subsection 3.1. Note that the probability density of $\mathrm{Dir}(\alpha_1, \alpha_2)$ diverges at the boundaries $\theta_1 = 0, 1$ if $\alpha_j < 1$. To ensure that our interpolation uses a density that favours values y_n that are closer to the expectation instead of favouring those that are close to the bounds y_i° and y_{i+1}°, we choose $\lambda \gg n_{\max}$ because this yields $\lambda(x_j^\circ - x_{j+1}^\circ) > 1$ with high probability. Therefore, we will later consider the limit $\lambda \to \infty$.

We now define the loss in each step as the Bayesian surprise:

Definition 5 (Bayesian loss function) *The loss in step n is defined by*

$$L_n^\lambda(y_n) := -\log \mathrm{pr}(y_n|x_1,\dots,x_n,y_1,\dots,y_{n-1})\,,$$

where the superscript λ reminds us that pr *already depends on λ.*

Due to

$$\mathrm{pr}(y_1,\dots,y_n|x_1,\dots,x_n) = \prod_{j=1}^{n} \mathrm{pr}(y_j|x_1,\dots,x_j,y_1,\dots,y_{j-1})$$

(just apply (17) for each j) we obtain:

Lemma 11 (loss until step $n_{\max}$) *The total loss for steps $1,\dots,n_{\max}$ in the prediction game reads:*

$$\sum_{n=1}^{n_{\max}} L_n^\lambda(y_n) \quad = \quad -\log \mathrm{pr}(y_1,\dots,y_{n_{\max}}|x_1,\dots,x_{n_{\max}})$$

The asymptotic for large λ of the total loss can be nicely described in terms of relative entropies:

Theorem 1 (asymptotic total loss)

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{n=1}^{n_{\max}} L_n^\lambda(y_n) = D(u\|f'_{n_{\max}})\,. \tag{19}$$

Hence,

$$\lim_{n_{\max}\to\infty} \left[\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{n=1}^{n_{\max}} L_n^\lambda(y_n)\right] = D(u\|f')\,. \tag{20}$$

Proof: To shorten notation, we write n and j for $n_{\max}$ and n, respectively. Taking the logarithm of (15) yields:

$$\log \mathrm{pr}(y_1,\dots,y_n|x_1,\dots,x_n) = \sum_{j=0}^{n} (\lambda\Delta_j^X - 1)\log\Delta_j^Y + \log\Gamma(\lambda) - \sum_{j=0}^{n} \log\Gamma(\lambda\Delta_j^X)\,. \tag{21}$$

We now use the Stirling approximation

$$\log\Gamma(z) = z\log z - z\log e + O(\log z)\,.$$

Thus,

$$-\sum_{j=0}^{n} \log\Gamma(\lambda\Delta_j^X) + \log\Gamma(\lambda) \quad = \quad -\lambda\sum_{j=0}^{n} \Delta_j^X \log\Delta_j^X - O(\log\lambda)\,.$$

Therefore,

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \log \mathrm{pr}(y_1,\dots,y_n|x_1,\dots,x_n) = \sum_{j=1}^{n} \Delta_j^X \log\Delta_j^Y/\Delta_j^X = -D(u\|f'_n)\,.$$

The second part of the statement holds because

$$\lim_{n\to\infty}\int_0^1 \log f_n'(x)dx = \int_0^1 \lim_{n\to\infty} \log f_n'(x)dx\,,$$

due to the bounded convergence theorem (the sequence $(\log f_n')_{n\in\mathbb{N}}$ is uniformly bounded because $\min_x\{f'(x)\} \leq f_n' \leq \max_x\{f'(x)\}$ by the mean value theorem). ■

We have seen that the complexity of the interpolation problem has turned out to depend on $D(u\|f')$ (for an appropriate limit, namely $\lambda \to \infty$). Since information geometry considers relative entropy as an analog of squared length in Euclidean geometry (Amari and Nagaoka, 1993), the total loss again depends on the squared length of the vector (u, g') or $(u, g^{-1'})$ in Figures 3(a) or 3(b), respectively, in analogy to Subsection 3.1 where it was given by $\mathrm{Var}(f')$ (i.e., the squared length of the vector g' or $g^{-1'}$ in Figures 2(a) or 2(b)).

4. Semi-Supervised Interpolation

In addition to the $n-1$ labelled points $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ and the unlabelled value x_n, we are now given the density p_X. For the anticausal scenario, i.e., if $X = E$ and $Y = C$, Lemmas 1 and 2 state positive correlation between p_X and f' or $\log f'$, respectively. Hence, large density $p(x)$ tends to correspond to large slope. Qualitatively, this already provides a guideline on how to modify the linear interpolation: the value x_n defines a partition of $(x_i^\circ, x_{i+1}^\circ)$ into two intervals. We first compare the average probability density in the left interval with the one in the right one. Whenever it is larger in the left one than in the right one, we slightly increase $\hat{y}_n$ because we expect the slope of f to be larger on the left interval. This, however, is just a rough intuition. The precise method of employing our knowledge on p_X depends on whether we use the correlations between f' and p_X or between $\log f'$ and p_X. We start with the former because the performance analysis of the corresponding SSL method uses a loss function that is closer to standard loss functions in machine learning.

4.1 SSL Using Correlations Between Slope and Density

In our SSL version, the estimation reads:

Definition 6 (additive SSL interpolation) *Let F denote the cumulative distribution of X and $s > 0$ be a parameter that controls how strongly the interpolation accounts for the distribution p_X. Then additive SSL interpolation is given by*

$$\hat{y}_n^s := \hat{y}_n + s\frac{(x_{i+1}^\circ - x_n)(x_i^\circ - x_n)}{x_{i+1}^\circ - x_i^\circ}\left[\frac{F(x_n) - F(x_i^\circ)}{x_n - x_i^\circ} - \frac{F(x_{i+1}^\circ) - F(x_n)}{x_{i+1}^\circ - x_n}\right]\,,$$

where $\hat{y}_n$ is defined as in (11). Note that s must be admissible in the sense that it is small enough to ensure that $\hat{y}_n^s$ remains inside the interval $(y_i^\circ, y_{i+1}^\circ)$.

To intuitively understand this interpolation, note that the term in the bracket is the difference between the average densities of the left and the right interval. Hence $\hat{y}_n^s$ is increased compared to the standard interpolation whenever the left interval contains higher density. Further understanding of why we define our SSL interpolation precisely in such a way will be provided below in the proof of Theorem 2. We first state our main result proved below:

Theorem 2 (total loss in terms of (co)-variances) *The total loss in the infinite interpolation game using $\hat{y}_n^s$ in Definition 6 reads:*

$$\sum_{n=1}^{\infty} L_n(y_n, \hat{y}_n^s) = \mathrm{Var}(f' - s\, p_X) = \mathrm{Var}(f') - 2s\,\mathrm{Cov}[f', p_X] + s^2\,\mathrm{Var}(p_X)\,.$$

In causal direction we have $\mathrm{Cov}[f', p_X] = 0$ and the additional term $s^2\,\mathrm{Var}(p_X)$ makes the performance worse than the baseline. In anticausal direction we have $\mathrm{Cov}[f', p_X] > 0$. Then standard linear regression tells us that the optimal improvement is reached for

$$s = \frac{\mathrm{Cov}[f', p_X]}{\mathrm{Var}(p_X)}\,,$$

if this value is admissible (otherwise one chooses a smaller one). Then the remaining loss reads:

$$\mathrm{Var}(f' - sp_X) \quad = \quad \mathrm{Var}(f') - \frac{(\mathrm{Cov}[f', p_X])^2}{\mathrm{Var}(p_X)}\,,$$

which is exactly the squared distance between v and g^{-1} in Figure 2(b). By Pythagoras, the squared length of $(v, 0)$ is the amount by which SSL improves the prediction for the optimal choice of s. We conclude:

Corollary 2 (Anticausal SSL works, causal SSL doesn't) *If $X = E$ and $Y = C$, SSL interpolation outperforms its supervised baseline version for sufficiently small s in the sense that*

$$\sum_{n=1}^{\infty} [L_n(y_n, \hat{y}_n^s) - L_n(y_n, \hat{y}_n)] < 0\,.$$

If $X = C$ and $Y = E$, SSL increases the total loss for all admissible s.

Finding the right value s needs to be a non-trivial problem for the following reason. p_E deviates from the uniform distribution for *two* reasons: first, because the function g is nonlinear and second, because p_C is not uniform. In other words, we do not know which part of the structure of p_E is due to the structure of g and which part due to the structure of p_C. This is also shown by the two extreme cases (1) where g is the identity and p_C and p_E are identical densities and (2) p_C is uniform and $p_E = g^{-1'}$. The optimal way to use p_E for better predicting g^{-1} will typically be a compromise that neither assumes that p_C is uniform nor that g is linear. The two extreme cases nicely correspond to a degeneration of the triangles in Figures 3(a) and 2(a): For linear g, the derivative $g^{-1'}$ is constant and thus coincides with the trivial random variable 0 and the trivial density u. On the other hand, for uniform p_C, $g^{-1'}$ and p_E coincide. For the generic case, the projection of $g^{-1'}$ onto the line from p_E to u is an interior point. Finding the right balance between attributing the structure of p_E entirely to the structure of g or entirely to the structure of p_C amounts to finding the projection points v and w_a that correspond to an optimal performance of our SSL methods in Subsections 4.1 and in Subsection 4.2, respectively. Since we do not know g^{-1}, we do not know the projection points v and w_a beforehand. Therefore, we have to work with the following heuristics: in step n, we choose the value s_{n-1} that minimizes the total loss until step $n-1$, which is easy to compute using Corollary 4 below.

The remainder of this subsection is devoted to the proof of Theorem 2 with some additional intuitive explanations at the end. To quantitatively analyze the loss, it is helpful to describe the estimation process as an estimation of the slope f'_n (which is equivalent) instead of an estimation of y_n. Let us define $\hat{f}_n$ as the function passing through $(x_n, \hat{y}_n)$ in addition to the points $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$. Standard linear interpolation obviously amounts to setting

$$\hat{f}'_n := f'_{n-1} \,.$$

Note that f'_{n-1} indicates the average slope for each open interval $(x^\circ_j, x^\circ_{j+1})$ and is undefined for each x°_j with $j = 0, \dots, n$. It is therefore convenient to consider f'_n as the following conditional expectation:

Lemma 12 (f'_n as conditional expectation of f') *Let $J_n : [0,1] \to \{0, \dots, n\}$ be the random variable such that for each x the value $J_n(x)$ indicates the subinterval in which x lies (defined by the n observed x-values $x^\circ_1, x^\circ_i, x_n, x^\circ_{i+1} \dots, x^\circ_{n-1}$). Then,*

$$f'_n = \mathbf{E}[f' | J_n] \,.$$

The proof is immediate via the mean value theorem. Similarly, we now introduce average densities:

Definition 7 (average density as conditional expectation) *Let J_n be defined as in Lemma 12. Then the average density (corresponding to the partition of $[0,1]$ defined by the first n x-values) is the function on $[0,1]$ given by*

$$p_n := \mathbf{E}[p_X | J_n] \,,$$

which is defined only in the interior of all $n+1$ intervals.

For $x \in (x^\circ_j, x^\circ_{j+1})$ with $j \neq i$ we have, for instance:

$$p_n(x) = \frac{F(x^\circ_{j+1}) - F(x^\circ_j)}{x^\circ_{j+1} - x^\circ_j} \,. \tag{22}$$

Using these conditional expectations, our SSL interpolation can be written in a concise form:

Lemma 13 (additive SSL interpolation in terms of conditional expectations) *The interpolation in Definition 6 amounts to setting*

$$(\hat{f}^s)'_n = f'_{n-1} + s(p_n - p_{n-1}) \,. \tag{23}$$

Proof: We only need to show that integrating (23) from x°_i to x_n yields the correct value for $\hat{y}^s_n$. On all intervals other than $(x^\circ_i, x^\circ_{i+1})$ (23) is certainly true because $\hat{f}^s_n$ coincides with f_{n-1} and $p_n - p_{n-1}$ is zero. On the interval (x°_i, x_n) the average densities p_{n-1} and p_n are given by

$$p_{n-1} = \frac{F(x^\circ_{i+1}) - F(x^\circ_i)}{x^\circ_{i+1} - x^\circ_i} \quad \text{and} \quad p_n = \frac{F(x_n) - F(x^\circ_i)}{x_n - x^\circ_i} \,.$$

Inserting this into (23) and integrating it from x_i° to x_n yields:

$$\begin{aligned}
\hat{y}_n^s &= \hat{y}_n + s\left[F(x_n) - F(x_i^\circ) - \frac{F(x_{i+1}^\circ) - F(x_i^\circ)}{x_{i+1}^\circ - x_i^\circ}(x_n - x_i^\circ)\right] \\
&= \hat{y}_n + s\left[\frac{F(x_n) - F(x_i^\circ)}{x_{i+1}^\circ - x_i^\circ}(x_{i+1}^\circ - x_i^\circ) - \frac{F(x_{i+1}^\circ) - F(x_i^\circ)}{x_{i+1}^\circ - x_i^\circ}(x_n - x_i^\circ)\right] \\
&= \hat{y}_n + \frac{s}{x_{i+1}^\circ - x_i^\circ}\left[-(x_n - x_i^\circ)F(x_{i+1}^\circ) + (x_{i+1}^\circ - x_i^\circ)F(x_n) - (x_{i+1}^\circ - x_n)F(x_i^\circ)\right] \\
&= \hat{y}_n + s\frac{(x_{i+1}^\circ - x_n)(x_i^\circ - x_n)}{x_{i+1}^\circ - x_i^\circ}\left[\frac{F(x_n) - F(x_i^\circ)}{x_n - x_i^\circ} - \frac{F(x_{i+1}^\circ) - F(x_n)}{x_{i+1}^\circ - x_n}\right].
\end{aligned}$$

■

Using Lemma 5 we are now able to phrase the loss in step n using our conditional expectations:

Corollary 3 (difference between interpolating functions) *The loss of the SSL version in step n reads*

$$L_n(y_n, \hat{y}_n^s) = \int_0^1 \left[(f'_n - f'_{n-1}) - s(p_n - p_{n-1})\right]^2 dx \tag{24}$$

$$= \int_0^1 [(f'_n - sp_n) - (f'_{n-1} - sp_{n-1})]^2 dx\,. \tag{25}$$

To derive a closed form for the total loss until step n we observe that f'_{n-1} and p_{n-1} can also be seen as conditional expectations of f'_n and p_n, respectively:

Lemma 14 (concatenating conditional expectations)

$$\mathbf{E}[f'|J_{n-1}] = \mathbf{E}[f'_n|J_{n-1}] \quad \textit{and} \quad \mathbf{E}[p_X|J_{n-1}] = \mathbf{E}[p_n|J_{n-1}]\,.$$

Proof: Applying the law of total expectation $\mathbf{E}[\mathbf{E}[A|B]] = \mathbf{E}[A]$ to each value of J_{n-1} yields $\mathbf{E}[\mathbf{E}[f'|J_n]|J_{n-1} = j] = \mathbf{E}[f'|J_{n-1} = j]$. Hence, $\mathbf{E}[\mathbf{E}[f'|J_n]|J_{n-1}] = \mathbf{E}[f'|J_{n-1}]$. The proof for p_X is similar. ■

Since we want to show that the total loss until step n can be written as a variance, we first need to rewrite the loss in each step as variance:

Lemma 15 (loss as variance of conditional expectation)

$$L_n(y_n, \hat{y}_n^s) = \mathbf{E}[\mathrm{Var}(f'_n - sp_n|J_{n-1})]\,.$$

Proof: The right-hand side of (24) can be written as

$$\begin{aligned}
&\int_0^1 ((f'_n - sp_n) - (f'_{n-1} - sp_{n-1}))^2 dx = \int_0^1 (f'_n - sp_n - \mathbf{E}[f' - sp_X|J_{n-1}])^2 dx \\
= &\int_0^1 (f'_n - sp_n - \mathbf{E}[f'_n - sp_n|J_{n-1}])^2 dx = \mathbf{E}[\mathrm{Var}(f'_n - sp_n|J_{n-1})]\,.
\end{aligned}$$

■

We can now express the total loss after n steps as a variance:

Lemma 16 (total loss after n steps)

$$\sum_{j=1}^{n} L_j(y_j, \hat{y}_j^s) = \mathrm{Var}(f'_n - sp_n)\,, \tag{26}$$

where the variance is again meant with respect to the Lebesgue measure.

Proof: By induction over n. Let (26) hold for n. Using the law of total variance we have

$$\begin{aligned} \mathrm{Var}(f'_n - sp_n) &= \mathbf{E}[\mathrm{Var}(f'_n - sp_n | J_{n-1})] + \mathrm{Var}(\mathbf{E}[f'_n - sp_n | J_{n-1}]) \\ &= L_n(\hat{y}_n^s, y_n) + \mathrm{Var}(f'_{n-1} - sp_{n-1})\,, \end{aligned}$$

where we have used Lemma 15. ■

As a simple conclusion we find:

Corollary 4 (optimal value s_n) *The total loss $\sum_{j=1}^{n} L_j(y_j, \hat{y}_j^s)$ until step n is minimized for*

$$s_n := \frac{\mathrm{Cov}[f'_n, p_n]}{\mathrm{Var}(p_n)}\,.$$

Moreover, s_n converges to the value s optimizing the total loss for infinitely many steps.

The limit $n \to \infty$ now proves Theorem 2:

$$\begin{aligned} \lim_{n\to\infty} \mathrm{Var}(f'_n - sp_n) &= \lim_{n\to\infty} \int_0^1 (f'_n - sp_n) - (1-s))^2 dx \\ &= \int_0^1 ((f' - sp_X) - (1-s))^2 dx = \mathrm{Var}(f' - sp_X)\,. \end{aligned}$$

Theorem 2 only states an improvement of the total loss over the infinite number of steps without stating for which n we get an improvement. The following remarks provide an intuition about in which steps SSL is effective. The term $\mathrm{Cov}[f'_n, p_n]$ quantifies to what extent the covariance of f' and p_X is apparent on the level of coarse-graining defined by the observations available in step n. For $n \to \infty$, it converges to $\mathrm{Cov}[f', p_X]$, which is positive in the anticausal scenario. The difference

$$\mathrm{Cov}[f'_n, p_n] - \mathrm{Cov}[f'_{n-1}, p_{n-1}] \tag{27}$$

measures to what extent the correlations between f' and p_X get better visible when the coarse-graining is made finer by going from $n-1$ to n intervals. One can easily show that (27) can be rewritten as $\mathrm{Cov}[f'_n - f'_{n-1}, p_n - p_{n-1}]$, which is positive whenever either (1) y_n is greater than the value $\hat{y}_n$ obtained by linear interpolation $\hat{y}_n := f_{n-1}(x_n)$ and the average probability density is larger on the left interval (x_i°, x_n) than on the right interval (x_n, x_{i+1}°) or (2) y_n is smaller than $\hat{y}_n$ and the density is larger on the right interval. Hence, (27) is positive whenever our SSL method corrects $\hat{y}_n$ in the correct direction. In other words, SSL does the right thing in step n whenever n defines a level of coarse-graining for which the covariance of f' and p_X gets better visible than in the previous step.

4.2 SSL Using Correlations Between Log Slope and Density

As in Subsection 4.1 we modify the interpolation in a way that favors functions that have higher derivative in regions where p_X is large. To do so, we use the Dirichlet process in Definition 4 with respect to a coordinate system that makes p_X more uniform: if we reparameterize X such that the differences Δ_j^X get larger in regions with high density, the interpolation with respect to the new coordinates infers the corresponding Δ_j^Y to be larger. To analyze the total loss for such a 'deformed interpolation' does not require to redo the computations in Subsection 3.2. Instead, we observe that applying the transformation $x_j \mapsto \tilde{x}_j = b(x_j)$ with some diffeomorphism b and performing interpolation in the new coordinate system amounts to interpolating $\tilde{f} := f \circ b^{-1}$. We thus conclude that the term on the right hand side of (20) is replaced with $D(u \| (f \circ b^{-1})')$. As an aside, we should mention that interpolation in the new coordinates amounts to using a Dirichlet process with a different base measure, namely the density that is uniform in the *new* coordinates. We conclude:

Lemma 17 (loss of deformed interpolation) *The asymptotic of the loss with respect to the above 'b-deformed interpolation' (denoted by $\tilde{L}$) reads:*

$$\lim_{n_{\max}\to\infty} \left[\lim_{\lambda\to\infty} \sum_{n=1}^{n_{\max}} \tilde{L}_n^\lambda(y_n) \right] = D(\tilde{u} \| f'), \tag{28}$$

where $\tilde{u} := b'$ denotes the density that is the image of the uniform under b^{-1}.

Proof: Since the density f' is the image of the uniform distribution under f^{-1}, the density $(f \circ b^{-1})'$ is the image of the uniform distribution under $b \circ f^{-1}$. Relative entropy is preserved under bijections, we can thus apply b^{-1} to the left argument u of $D(.\|.)$ (which generates the density b') instead of applying b to the right one. ■

We can now easily compare the performance of interpolations with respect to different coordinate systems:

Lemma 18 (comparing Dirichlet interpolations)

$$\lim_{n_{\max}\to\infty} \left[\lim_{\lambda\to\infty} \sum_{n=1}^{n_{\max}} (L_n^\lambda(y_n) - \tilde{L}_n^\lambda(y_n)) \right] = D(u\|f') - D(\tilde{u}\|f') \,.$$

Given the relation between performance and the relative entropy stated by Lemma 18 we conclude:

Corollary 5 (benefit of changing the coordinate system) *The deformed interpolation with respect to a transformation that turns $\tilde{u}$ into the uniform distribution on* $[0,1]$ *asymptotically outperforms the standard interpolation for $n \to \infty$ if and only if*

$$D(\tilde{u}\|f') < D(u\|f') \,.$$

We now define the density that generates our SSL interpolation:

Definition 8 (SSL interpolation) *Let w_s be the mixture of p_X with the uniform distribution, i.e.,*

$$w_s = sp_X + (1-s)\,.$$

Apply the coordinate transformation that transforms w_s into the uniform distribution, i.e.,

$$W_s(x) := sF(x) + (1-s)x\,. \tag{29}$$

Then the deformed interpolation is our usual Dirichlet interpolation from Subsection 2.2 applied to the values $\tilde{x}_j := W_s(x_j)$.

We then state our main result regarding the performance of SSL by Dirichlet process in the modified coordinate system:

Theorem 3 (improvement of performace by SSL) *Predicting y_n via the Dirichlet process in the coordinate system W_s, as defined by (29), improves the performance by the amount $D(u\|w_s)$.*

To further understand our deformed interpolation one may wonder whether the expectation of y_n coincides with the value $\hat{y}_n^s$ in Subsection 4.1. Remarkably, this is not the case. Instead, it turns out that the SSL method in this subsection modifies the slope by a *multiplicative factor* that accounts for p_X while the SSL method in Subsection 4.1 corrects the slope by an *additive summand*. This nicely corresponds to the fact that Subsection 4.1 employs correlations between p_X and f' while this Subsection employs correlations between p_X and $\log f'$. This difference is made more explicit in Appendix D.

We now state our main result:

Theorem 4 (anticausal SSL works, causal SSL doesn't) *Let cause C and effect E satisfy Assumption 2. For $X := E$ and $Y := C$ and $f := g^{-1}$ there is an $s > 0$ for which the deformed interpolation outperforms standard linear interpolation. For $X := C$ and $Y := E$ and $f := g$, there is no such s.*

Proof In the terminology of information geometry (Amari and Nagaoka, 1993; Amari, 2001), $M := \{w_s\}_{s\in I}$ is an m-manifold. There is therefore a unique minimizer w_a of the distance $D(w_s\|f')$ (called the 'projection' of f' onto M) satisfying the orthogonality, see Eq. (60) in (Amari, 2001),

$$D(u\|f') = D(u\|w_a) + D(w_a\|f')\,. \tag{30}$$

For $X = C$ and $Y = E$, we have $w_a = u$. Therefore, M cannot contain any w_s for which $D(w_s\|f') < D(u\|f')$.

For the causal scenario $X = E$ and $Y = C$, we consider the function

$$h(s) := D(w_s\|f') = \int_0^1 (sp(x) - (1-s)) \log \frac{sp(x) - (1-s)}{f'(x)} dx\,. \tag{31}$$

Its derivative reads

$$h'(s) = \int_0^1 (p(x) - 1) \log \frac{w_s(x)}{f'(x)} dx\,. \tag{32}$$

We observe

$$h'(0) = \int_0^1 (p(x) - 1) \log \frac{1}{f'(x)} dx = -\mathrm{Cov}[p_X, \log f'] \,.$$

Using (6), we thus have $h'(0) < 0$. Therefore, the unique minimum a of h satisfies $a > 0$. ■

Remark 1 *We show in Appendix C that $a \leq 1$ holds in addition to $a > 0$ whenever one assumes the additional independence postulate*

$$\mathrm{Cov}[g', \log p_E] = 0 \,,$$

which has not been described in the literature yet.[5] *Then w_a is a* mixture *of u and p_X.*

In strong analogy to Subsection 4.1, the theory does not tell us how to find the optimal value $s = a$. We know that w_a is geometrically given by projecting f' onto the line connecting u and p_X, see Figure 3(b), but since we don't know f', it is not even clear how to find *any* w_s that is closer to f' than u is. In Subsection 4.1 we have provided intuitive arguments why this needs to be a non-trivial problem: The free parameter s defines a prior decision to what extent we attribute the non-uniformness of p_X to p_Y and to what extent to the non-linearity of f. Again, we propose the following heuristic procedure to iteratively adapt s during the SSL procedure: in each step n, we already know which value a_{n-1} minimizes $D(w_s \| f'_{n-1})$. In other words, among all possible deformations given by w_s, we can choose the one that yields the best prediction for the piecewise linear function f'_{n-1} interpolating the known values. Then, a_n converges to the optimal value a as shown by the following result which is proved in Appendix B:

Lemma 19 (continuity of projections) *Let f' be continuous and p_X be bounded from above. Define*

$$a_n := \mathrm{argmin}_{s \in I} D(w_s \| f'_n) \,.$$

Then we have

$$\lim_{n \to \infty} a_n = \mathrm{argmin}_{s \in I} D(w_s \| f') \,.$$

5. Conclusions

We have analyzed a semi-supervised interpolation for $Y = f(X)$ for an unknown strictly monotonically increasing function f. Whenever Y is the cause and X the effect the derivative of f tends to be high in regions where p_X is large – provided that one believes in the model assumptions of Information-Geometric Causal Inference. We have proposed two different SSL methods, one employs the fact that p_X is positively correlated with f', while the other one employs positive correlations between p_X and the logarithm of the slope. In both cases, the SSL method changes the value $\hat{y}_n$ inferred by standard linear interpolation by an amount that depends on the average probability densities of X in the intervals between x_n and the closest point to the left and to the right. It turns out that such a modified linear

[5] It turns out to be equivalent to the dual version of (7) by replacing each relative entropy $D(p\|q)$ with $D(q\|p)$. It is known in information geometry (Amari, 2001) that many theorems have such a 'dual' counterpart.

interpolation outperforms standard linear interpolation with respect to two substantially different loss functions: the first one is a squared distance, the second one the Bayesian surprise.

To the best of our knowledge, this is the first theoretical result that links the performance of SSL to the causal direction, provided that one accepts the underlying independence assumption for P_{cause} and $P_{\text{effect}|\text{cause}}$. SSL-algorithms that employ P_X by changing the geometry of the input space accordingly have been described earlier Chapelle et al. (2006). For instance, P_X may define a notion of smoothness (e.g. via P_X-dependent kernels or graphs) and thus influence the regularization term. Here we have justified an appropriate change of the geometry based on a postulate that is linked to the causal direction.

Certainly, the notion of (in)dependence of P_X and $P_{Y|X}$ used throughout this article is a rather simplistic one. First, the deterministic scenario applies only to very specific causal relations in real life. Second, even for this case, one would not expect that independence between P_{cause} and $P_{\text{effect}|\text{cause}}$ always holds in the sense of vanishing correlations as discussed here. To find notions of (in)dependence that turn out to be related to the causal direction in realistic learning scenarios has to be left to the future.

Acknowledgments

The authors would like to thank Eleni Sgouritsa, Joris Mooij, and Jonas Peters for helpful remarks on the manuscript.

Appendix A. Proof of Lemma 10

Using (17) yields

$$\begin{aligned} & \log \mathrm{pr}(y_n|x_1,\dots,x_n,y_1,\dots,y_{n-1}) \\ = \; & \mathrm{pr}(y_1,\dots,y_n|x_1,\dots,x_n) - \log \mathrm{pr}(y_1,\dots,y_{n-1}|x_1,\dots,x_{n-1})\,. \end{aligned}$$

We now compare the terms in (21) with those that occur in the same formula for $n+1$: the term

$$(\lambda(x^\circ_{i+1}-x^\circ_i)-1)\log(y^\circ_{i+1}-y^\circ_i) \tag{33}$$

is replaced with

$$(\lambda(x_{n+1}-x^\circ_i)-1)\log(y_{n+1}-y^\circ_i)+(\lambda(x^\circ_{i+1}-x_n)-1)\log(y^\circ_{i+1}-y_n)\,. \tag{34}$$

Splitting the term (33) into

$$(\lambda(x_n-x^\circ_i)-1)\log(y^\circ_{i+1}-y^\circ_i)+(\lambda(x_n-x^\circ_{i+1})-1)\log(y^\circ_{i+1}-y^\circ_i)+\log(y^\circ_{i+1}-y^\circ_i)\,,$$

the difference between (33) and (34) can be written as

$$\lambda(x_n-x^\circ_i)-1)\log\frac{y_n-y^\circ_i}{y^\circ_{i+1}-y^\circ_i}+\lambda(x^\circ_{i+1}-x_n)-1)\log\frac{y^\circ_{i+1}-y_n}{y^\circ_{i+1}-y^\circ_i}-\log(y^\circ_{i+1}-y^\circ_i)\,.$$

To understand how the normalization factors change from $n-1$ to n we observe that the term

$$\log\Gamma(\lambda(x^\circ_{i+1}-x^\circ_i))$$

is replaced with

$$\log \Gamma(\lambda(x_n - x_i^\circ)) + \log \Gamma(\lambda(x_{i+1}^\circ - x_n)) .$$

Then the statement follows.

Appendix B. Proof of Lemma 19

We first consider the affine family of densities $q_\lambda := \lambda r_1 + (1-\lambda) r_2$, where r_1, r_2 are strictly positive densities with non-zero lower bound b. We then show that the value s minimizing $D(w_s \| q_\lambda)$ depends continuously on λ. To this end, we introduce the function

$$\ell(\lambda, s) := \frac{d}{ds} D(w_s \| q_\lambda) = \int (p(x) - 1) \log \frac{w_s(x)}{q_\lambda} dx ,$$

where the last equality is derived in analogy to (32) by replacing f' with q_λ in. Then

$$\begin{aligned} \frac{\partial}{\partial \lambda} \ell(\lambda, s) &= -\int_0^1 \frac{p(x) - 1}{\lambda r_1(x) + (1-\lambda) r_2(x)} (r_1(x) - r_2(x)) dx \\ \frac{\partial}{\partial s} \ell(\lambda, s) &= \int_0^1 \frac{(p(x)-1)^2}{w_s(x)} dx . \end{aligned}$$

Let $b > 0$ be a lower bound for r_1 and r_2 and $d > 0$ an upper bound for p_X. We then obtain

$$\left| \frac{\partial}{\partial \lambda} \ell(\lambda, s) \right| \leq \frac{d}{b} \int |r_1(x) - r_2(x)| dx . \tag{35}$$

Moreover,

$$\left| \frac{\partial}{\partial s} \ell(\lambda, s) \right| \geq \frac{1}{1+d} \int (1 - p(x))^2 dx . \tag{36}$$

Since (36) is non-zero because p_X is not the constant function 1 (otherwise it could not correlate with $\log f'$), the law of implicit functions states that we can locally find a function v (around some solution a) by

$$\ell(\lambda, v(\lambda)) = 0 ,$$

with

$$v'(\lambda) = \frac{\partial}{\partial \lambda} \ell(\lambda, a) \left(\frac{\partial}{\partial s} \ell(\lambda, a) \right)^{-1} .$$

The difference between the s-values s_1 and s_2 for r_1 and r_2, respectively, can be bounded from above by

$$|v(1) - v(0)| \leq \sup_{\lambda \in [0,1]} |v'(\lambda)| \leq \frac{d(d+1)}{b} \frac{\int |r_1(x) - r_2(x)| dx}{\int (1 - p(x))^2 dx} , \tag{37}$$

where the last inequality follows from combining (35) and (36). Since each f_n' is strictly positive and f_n' converges uniformly to f', which is strictly positive on the compact interval $[0,1]$, we can find a uniform lower bound b for the functions f_n'. Using (37) with $r_1 := f_n'$ and $r_2 := r_2$ yields

$$|a_n - a| \leq \frac{d(d+1)}{b} \int |f_n'(x) - f'(x)| dx .$$

Then the right hand side converges to zero, again due to the uniform convergence of f_n' to f'.

Appendix C. Using the Dual Independence Postulate

Straightforward computation shows that the 'dual' independence postulate

$$\mathrm{Cov}[g', \log p_E] = 0$$

is equivalent to

$$D(g' \| p_C) = D(g' \| u) + (u \| p_C)\,.$$

Applying the function g to all distributions yields

$$D(u \| p_E) = D(u \| g^{-1'}) + D(g^{-1'} \| p_E)\,. \tag{38}$$

For the function h defined in (31) we observe that

$$\begin{aligned} h'(1) &= \int_0^1 (p(x) - u(x)) \log \frac{p(x)}{f'(x)} dx \\ &= D(p_X \| f') + D(u \| p_X) - D(u \| f')\,. \end{aligned}$$

Using

$$D(u \| p_X) = D(u \| f') + D(f' \| p_X)\,,$$

due to (38) yields

$$h'(1) = D(p_X \| f') + D(f' \| p_X) \geq 0\,,$$

with equality only for $f' = p_X$, i.e., if p_Y is uniform. Therefore the unique minimum a of h satisfies $s \leq 1$ with equality only for uniform input.

Appendix D. Comparing the Two Interpolation Schemes

We now explain why the SSL interpolation in Subsection 4.2 differs from the one in Subsection 4.1 not only by the fact that the former infers one specific value $\hat{y}_n^s$ while the latter provides a conditional distribution. We now see that the expectation of the conditional of the SSL version of the Dirichlet process does not coincide with $\hat{y}_n^s$ in Subsection 4.1. To this end, we recall that the expectation for the standard linear interpolation in Subsection 3.1 reads

$$\hat{y}_n = \frac{x_n - x_i^\circ}{x_{i+1}^\circ - x_i^\circ}(y_{i+1}^\circ - y_i^\circ) + y_i^\circ\,.$$

Now, we just have to replace each x-value by $W_s(x)$ and obtain:

Lemma 20 (expectation of deformed interpolation)

$$\hat{y}_n^s = \frac{W_s(x_n) - W_s(x_i^\circ)}{W_s(x_{i+1}^\circ) - W_s(x_i^\circ)}(y_{i+1}^\circ - y_i^\circ) + y_i^\circ\,. \tag{39}$$

To understand (39), we note that it amounts to multiplying the slope of f_n with some factor:

Lemma 21 (deformed interpolation in terms of derivatives)

$$(\hat{f}_n^s)' = f'_{n-1} \frac{(1-s) + sp_n}{(1-s) + sp_{n-1}} = f'_{n-1} \frac{w_n^s}{w_{n-1}^s}\,, \tag{40}$$

with $w_n^s := \mathbf{E}[w_s | J_n]$.

Proof: Rewrite (39) as

$$\frac{\hat{y}_n^s - y_i^\circ}{x_n - x_i^\circ} = \frac{y_{i+1}^\circ - y_i^\circ}{x_{i+1}^\circ - x_i^\circ} \frac{W_s(x_n) - W_s(x_i^\circ)}{x_n - x_i^\circ} \frac{x_{i+1}^\circ - x_i^\circ}{W_s(x_{i+1}^\circ) - W_s(x_i^\circ)}\,.$$

Then the right hand side can be written as $f'_n w_n^s / w_{n-1}^s$. ■

To compare (40) to (23) we observe

$$s(p_n - p_{n-1}) = w_n^s - w_{n-1}^s\,.$$

Hence, (23) can also be written as

$$(\hat{f}_n^s)' = f'_{n-1} + w_n^s - w_{n-1}^s,\,.$$

Therefore the additively deformed interpolation modifies f'_{n-1} by adding the difference $w_n^s - w_{n-1}^s$ as summand, while the SSL interpolation in Subsection 4.2 is multiplicative in the sense that it adds the quotient w_n^s / w_{n-1}^s as factor.

References

S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

S. Amari and H. Nagaoka. *Methods of Information Geometry.* Oxford University Press, 1993.

N. Balakrishnan and Nevzorov V. *A Primer on Statistical Distributions.* John Wileys, New Jersey, USA, 2003.

S. Ben-David, T. Lu, and D. Pl. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In R. Servedio and T. Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 33–44. Omnipress, 2008.

D. Blackwell. Discreteness of Ferguson selections. *The Annals of Statistics*, 1(2):356–358, 1973.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, USA, 2006.

P. Daniusis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 143–150. AUAI Press, 2010.

M. Darnstädt, H. Simon, and B. Szörényi. Unlabeled data does provably help. In N. Portier and T. Wilke, editors, *30th International Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 20 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 185–196, 2013.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182–183:1–31, 2012.

D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. Justifying information-geometric causal inference. In V. Vovk, H. Papadopolous, and A. Gammerman, editors, *Measures of Complexity*, Festschrift for Alexey Chervonencis, pages 253–265. Springer Verlag, Heidelberg, 2015.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In Langford J. and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262. ACM, 2012.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. Semi-supervised learning in causal and anticausal settings. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference*, Festschrift in Honor of Vladimir Vapnik, pages 129–141. Springer, 2013.

E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. In G. Lebanon and S. Vishwanathan, editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings, 2015.

G. Yuanyuan, X. Niu, and H. Zhang. An extensive empirical study on semi-supervised learning. In G. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *10th IEEE International Conference on Data Mining (ICDM)*, pages 186–195. IEEE Computer Society, 2010.

T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In P. Langley, editor, *17th International Conference on Machine Learning (ICML)*, pages 1191–1198. Morgan Kaufmann Publishers, 2000.