On the Interventional Kullback-Leibler Divergence

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Modern machine learning approaches excel in static settings where a large amount of i.i.d. training data are available for a given task. In a dynamic environment, though, an intelligent agent needs to be able to transfer knowledge and re-use learned components across domains. It has been argued that this may be possible through causal models, aiming to mirror the modularity of the real world in terms of independent causal mechanisms. However, the true causal structure underlying a given set of data is generally not identifiable, so it is desirable to have means to quantify differences between models (e.g., between the ground truth and an estimate), on both the observational and interventional level.

In the present work, we introduce the Interventional Kullback-Leibler (IKL) divergence to quantify both structural and distributional differences between models based on a finite set of multi-environment distributions generated by interventions from the ground truth. Since we generally cannot quantify all differences between causal models for every finite set of interventional distributions, we propose a sufficient condition on the intervention targets to identify subsets of observed variables on which the models provably agree or disagree.

Keywords: causal distance, causal discovery, multi-environment learning

1. Introduction

Classical machine learning methods are concerned with learning a distribution P (or properties thereof) from a large i.i.d. sample using a simpler model Q. Recent advances in causality have shown how access to heterogeneous data from multiple environments can help uncover causal relationships within the data and address problems like adversarial robustness or domain generalization (Peters et al., 2017; Zhang et al., 2013; von Kügelgen et al., 2021; Schölkopf et al., 2021; Eastwood et al., 2022; Guo et al., 2022). In this framework, distributional shifts emerge from interventions in some underlying causal model $\mathscr{P}$. Identifying the true causal model, however, is difficult even in the multi-environment regime, requiring many environments and (often unverifiable) assumptions on the type of interventions present. One thus often needs to be satisfied with an approximate model $\mathscr{Q}$, which is *close* to a hypothetical ground truth $\mathscr{P}$.

In the present work, we propose a quantitative notion of observational and interventional *closeness* between two causal models $\mathscr{P}$ and $\mathscr{Q}$. Extending the classical Kullback-Leibler divergence, our *Interventional Kullback-Leibler divergence* $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})$ is computed between a finite set of interventional distributions $(P^e)_{e \in \mathcal{E}}$ from $\mathscr{P}$, denoted as $\mathscr{P}_{\mathcal{E}}$ and our model estimate $\mathscr{Q}$, consisting of a single reference distribution estimate Q and graphical model estimate G_Q. Intuitively, we can understand D_{IKL} as quantifying the average deviation between our model's predictions and distributions under interventions had the interventions also happened on our model estimate. We thus quantify how good our model estimate is beyond observational. Minimizing the IKL divergence over our estimate $\mathscr{Q}$ with respect to arbitrary interventions is equivalent to finding the true causal model $\mathscr{P} = \mathscr{Q}$. In practice, having access to only a limited amount of multi-environment distributions presents interesting challenges regarding which structural differences are identifiable. We

© 2023 .

propose a sufficient condition on the intervention targets trading-off a priori knowledge about $\mathscr{P}$ to establish the equivalence $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0 \iff \mathscr{P} = \mathscr{Q}$. If this condition is not satisfied, we further show which differences with respect to subsets of variables can be provably identified.

Previous approaches (Acid and Campos, 2003; Tsamardinos et al., 2006; Peters and Bühlmann, 2013) define interventional distances via differences between two causal graphs. In contrast, our IKL divergence measure (1) does not require access to the true causal graph G_P and (2) also takes into account distributional differences. Further, unlike Acharya et al. (2018); Peyrard and West (2020), there is no need to experimentally perform (hard-)interventions for the evaluation of our distance; instead we allow for arbitrary unknown mechanism changes, as long as the targets of the interventions are discerned.

The remainder of this work is organized as follows: In Section 2, we introduce the problem settings and the assumptions to set the stage for the following theoretical analysis. Section 3 contains a discussion about related work and our main results, studying the Interventional Kullback-Leibler divergence first for the example of a known graph and then in the general setting. Finally, in Section 4, we discuss the operational significance of the IKL divergence and various applications related to the area of multi-environment causal learning.

- We introduce the Interventional Kullback-Leibler divergence D_{IKL} that quantifies distributional and structural differences between a given set of multi-environment distributions $(P^e)_{e \in \mathcal{E}}$ from the ground truth $\mathscr{P}$ and our inferred model $\mathscr{Q}$.

- We prove a theorem establishing equivalence between $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ and equality between causal models $\mathscr{P} = \mathscr{Q}$ under certain conditions on the environments. We thereby trade off (partial) knowledge about $\mathscr{P}$ and access to interventional distributions $(P^e)_{e \in \mathcal{E}}$. This gives rise to a sufficient condition that can be verified using only the inferred model $\mathscr{Q}$.

2. Problem Setting and Preliminaries

We begin by reviewing the causal formalism used below. This is based on interventional distribution shifts, modelled as causal graphical models (Pearl, 2009).

2.1. Causal Terminology

Definition 1 (Causal Graphical Model) *A Causal Graphical Model (CGM) (P, G_P) over a variable set $\mathbf{X} = \{X_1, \dots, X_d\}$ consists of a distribution P and a directed acyclic graph (DAG) $G_P = (\mathbf{V}, E)$ where $\mathbf{V} := \{1, \dots, d\}$ is an index set and $\mathbf{E} \subseteq \mathbf{V}^2$ is an edge set, such that $(i, j) \in \mathbf{E}$ if and only if X_i directly causes X_j. We say that two CGMs $(P, G_P), (Q, G_Q)$ are the same $(P, G_P) = (Q, G_Q)$ if and only if $P = Q$ and $G_P = G_Q$.*

Definition 2 (Markov Factorization) *Given a joint distribution P and a DAG G, we say P satisfies the Markov factorization property[1] with respect to G (or is Markovian with respect to G) if*

$$P(X_1, \dots, X_d) = \prod_i P(X_i \mid \mathbf{PA}_i^G), \tag{1}$$

where $\mathbf{PA}_i^G$ denotes the set of parents of X_i in G.

1. We assume that all distributions have densities with respect to the Lebesgue measure.

While Definition 2 allows us to deduce properties of the distribution P from the corresponding graph G_P, the following faithfulness property allows us to perform inference in the converse direction, i.e. finding the structure of the graph given the distribution.

Definition 3 (Causal Faithfulness) *A joint distribution P is called faithful with respect to a DAG G if any conditional independence relationship in P is implied by d-separation in G (Spirtes et al., 1993; Peters et al., 2017).*

Assumption 4 *In this paper, we consider the set of distributions that are both Markovian and faithful with respect to some causal DAG G, i.e. given an observed distribution P, there exists a compatible graph G_P.*

A justification for the faithfulness assumption is given by Koller and Friedman (2010, Theorem 3.5) stating that almost all distributions P that are Markovian with respect to G are also faithful to G.

2.2. Multi-environment learning

Throughout this work, we take advantage of data from multiple environments. Specifically, we assume that given a set of multi-environment distributions $(P^e)_{e\in\mathcal{E}}$ over some variable set $\mathbf{X}$, there exists an underlying ground truth CGM $\mathscr{P} := (P, G_P)$ giving rise to $(P^e)_{e\in\mathcal{E}}$. The generative process behind these distributions is assumed to satisfy the below (Schölkopf et al., 2012; Peters et al., 2017):

Assumption 5 (Independent Causal Mechanisms (ICM)) *A causal generative process of a system of variables is composed of autonomous modules that do not inform or influence each other. Mathematically speaking, given a Markov factorization of the joint distribution as in (1), the causal conditionals (also called* mechanisms*) $P(X_i \mid \mathbf{PA}_i^{G_P})$ represent autonomous modules in the sense that*

(i) intervening on one $P(X_i \mid \mathbf{PA}_i^{G_P})$ will not change other mechanisms $P(X_j \mid \mathbf{PA}_j^{G_P}), i \neq j$,

(ii) knowing one mechanism will not provide information about other mechanisms.

Under Assumption 5, environment shifts act independently on each of the mechanisms $P(X_i \mid \mathbf{PA}_i^{G_P})$. Thus, for each environment, $e \in \mathcal{E}$ there exists a well-defined set of mechanisms that are altered by the environment.

Assumption 6 (Multi-Environment Distributions) *For each environment $e \in \mathcal{E}$, 'Nature' first chooses a subset of variables to intervene upon $\mathcal{I}_e \subseteq [d]$, where $[d] := \{1, \dots, d\}$. For each intervened variable, 'Nature' then chooses a mechanism shift $\tilde{P}(X_i \mid \mathbf{PA}_i) \in \mathcal{F}$ to perform, where $\mathcal{F}$ is a family of functions. The transition from P to P^e thus maps $P(X_i \mid \mathbf{PA}_i) \mapsto \tilde{P}(X_i \mid \mathbf{PA}_i)$ if $i \in \mathcal{I}_e$ and $P(X_i \mid \mathbf{PA}_i) \mapsto P(X_i \mid \mathbf{PA}_i)$ otherwise, and we may represent the joint distribution in environment e as:*

$$P^e(X_1, \dots, X_n) = \prod_{i\in\mathcal{I}^e} \tilde{P}(X_i \mid \mathbf{PA}_i) \cdot \prod_{i\in[d]\setminus\mathcal{I}^e} P(X_i \mid \mathbf{PA}_i) \tag{2}$$

Such mechanism shifts arise when the context of the observed distribution changes. This may happen for example as the result of different outer circumstances like climate or time (Mooij et al., 2016) or by explicitly intervening on an experiment. Such shifts are generally called *interventions*. Special cases include *soft interventions* where the structure of the graph is left invariant, i.e., X_i does not become conditionally independent of any of its parents, and *hard interventions* where $\tilde{P}(X_i \mid \mathbf{PA}_i) = \delta(X_i - x)$. In any case, P^e will still be Markovian with respect to the ground truth graph G_P. However, it may no longer be faithful to it, unless the interventions are soft.

Our final assumption concerns the existence of hidden confounder variables. We here relax the causal sufficiency assumption (stating that there are no unobserved confounders) to pseudo causal sufficiency. This ensures in particular that there are no changes in the observed mechanisms that are not explained by the environment.

Assumption 7 (Pseudo causal sufficiency) *The effect of any unobserved confounder can be completely explained by the environment variable, i.e. in any given environment $e \in \mathcal{E}$ potential unobserved confounders are fixed, and causal sufficiency is satisfied (Huang et al., 2019).*

3. The Interventional KL divergence

In this section, we introduce our proposed Interventional KL divergence. We begin by reviewing the problem of defining a distance between two CGMs $\mathscr{P} = (P, G_P)$ and $\mathscr{Q} = (Q, G_Q)$ and discussing related work. In the following sections we explain how distributional differences between P and Q and structural differences between G_P, G_Q are encoded in the IKL divergence respectively. Thereby, $\mathscr{Q}$ takes the role of an 'estimate' for the unknown $\mathscr{P}$, i.e. we know G_Q and have access to and can evaluate the density of Q. For complete proofs of the results in this section, we refer to Appendix A.

3.1. Related Work

Defining a general distance between two causal graphical models $(P, G_P), (Q, G_Q)$ requires comparing them both with respect to distributional differences between P, Q and structural differences between their underlying causal graphs G_P, G_Q. Early work addressing this problem considered only structural differences. The Structural Hamming Distance (SHD) (Acid and Campos, 2003; Tsamardinos et al., 2006) counts the number of different edges between two graphs, which has been found to provide limited insights about the effect on interventional distributions (Peters and Bühlmann, 2013). They instead proposed the Structural Intervention Distance (SID) to address this limitation by counting instead the number of different interventional distributions induced by two causal graphs. If the true causal graph G_P is unknown, it is generally unclear how to define a valid distance measure between the causal models. One avenue to solve this problem is leveraging multi-environment distributions that under Assumptions 6 and 7 can provide insights about the structural properties of the true causal graph. Peyrard and West (2020); Acharya et al. (2018) propose a score for comparing general causal graphical models based on their distributional distance after performing hard interventions on both of them. In section 3.3, we show that under these assumptions, the Interventional KL divergence entails their distance as a special case.

Another related area of research is that of multi-environment causal discovery, which seeks to uncover the true causal graph G_P by relaxing the i.i.d. assumption to exchangeable data (Guo et al., 2022) or interventional data (Eaton and Murphy, 2007; Ghassami et al., 2018; He and Geng, 2016).

Since these methods usually only provide probabilistic guarantees about recovering the true causal graph, a causal distance metric can be seen as an evaluation tool to track their progress.

Finally, there is a line of related work discussing the identifiability of causal effects and interventional Markov equivalence classes (Tian and Pearl, 2002; Hauser and Bühlmann, 2011, 2015). In this work, we are more interested in quantifying differences beyond Markov equivalence classes, assuming that these can be obtained from observational data.

3.2. Distributional Differences in the IKL divergence

Before defining the Interventional KL divergence for general causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ in the next section, we start by outlining how distributional differences between P and Q alone are encoded in the IKL divergence. To this end, we assume that the true causal graph is shared and known, i.e. $G_P = G_Q$. Throughout this work, we use the Kullback-Leibler divergence for quantifying distributional differences between P and Q, assuming that P is absolutely continuous with respect to Q:

$$D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}, \tag{3}$$

It is known that $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = 0$ if and only if $P = Q$. Now suppose that P^e arises from P via a mechanism shift on the variables $\mathbf{X}_{\mathcal{I}_e}$ for a subset $\mathcal{I}_e \subseteq [d]$. From the chain rule for relative entropy (see e.g. (Cover and Thomas, 2006, Thm 2.5.3)), we can deduce the following decomposition to understand how the effect of the mechanism shifts shows up in $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X}))$. From here on, we denote $\mathbb{E}_{x \sim P(x)}[X]$ as $\mathbb{E}_{P(x)}[X]$.

Lemma 8 *Let P, Q be distributions sharing the same causal graph G_P and assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = & \sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}))] \\ & + \underbrace{\sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}))]}_{=: D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}} \| \mathscr{Q})} \end{aligned} \tag{4}$$

As a special case we have for $\mathcal{I}_e = \emptyset$:

$$D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \sum_{i \in [d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}))] \tag{5}$$

The first sum in equation (4) goes over the intervened mechanisms and therefore fully characterizes the changes that are directly due to environment shifts. Note that it vanishes if we could replicate the environment shift on our model estimate Q. The second sum, which is a special case of the Interventional KL divergence to be defined more generally below, goes over the non-intervened variables, quantifying the distributions' distance resulting downstream from intervened variables due to the fact that the expectations over parents are taken with respect to intervened distributions.

Now assume that P is unobserved, but we still want to understand whether Q is a valid model. Since, P, Q are Markovian with respect to the same DAG G_P, we can express agreement between P and Q equivalently in terms of vanishing Interventional KL divergence $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})$.

First suppose that $P = Q$. The first sum in equation (4) can become arbitrarily large, inflating the KL divergence $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X}))$. However, the second sum in equation (4) vanishes, i.e., $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, since agreement between the conditional distributions is irrespective of the parent distributions taken in the expectation.[2] The IKL divergence therefore discounts the additional difference between P^e and Q introduced by the environment shift.

Conversely assume that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$. It is clear that this does not necessarily imply that $P = Q$, since we only sum over a subset of the local divergences composing the difference between the joint distribution P and Q, as shown in equation (5). But given multiple interventional distributions $(P^e)_{e \in \mathcal{E}}$, such that each mechanism of $P(\mathbf{X})$ remains unchanged in (at least) one environment, we can obtain a valid distance measure between P and Q via averaging. These observations are formalized in the following definition.

Definition 9 (Interventional KL divergence for shared causal graphs) *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_P)$ be two causal models sharing the same causal graph. Further, assume that we have access to a set of interventional distributions $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$ generated from $\mathscr{P}$. We define the Interventional KL divergence as*

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})} [D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_P}) \| Q(X_i \mid \mathbf{pa}_i^{G_P}))] \tag{6}$$

If $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, we say that our model $\mathscr{Q}$ is interventionally equivalent with $\mathscr{P}$ with respect to $\mathcal{P}$, denoted as $\mathscr{P} \approx_{\mathcal{E}} \mathscr{Q}$

Remark 10 *Under Assumptions 6, 7 and that no variable is always intervened upon, i.e. $\bigcap_{e \in \mathcal{E}} \mathcal{I}_e = \emptyset$, we can conclude that $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \approx_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.*

Note that if the reference distribution $P(\mathbf{X})$ is included in our set of interventional distributions, i.e. there exists $e \in \mathcal{E}$ such that $\mathcal{I}_e = \emptyset$, the KL divergence $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X}))$ makes up one of the terms in the IKL divergence. Thus, the equivalence $\mathscr{P} \approx_{\mathcal{E}} \mathscr{Q} \iff \mathscr{P} = \mathscr{Q}$ is trivially satisfied. We now illustrate an application of the case where we do not have access to the reference distribution P of the CGM (P, G_P).

Example 11 (Partial Observability) *In this example, we illustrate how the ideas presented above formalize the common conception of recovering the joint distribution $P(X_1, X_2, X_3)$ without observing all variables at once — given that we know its underlying DAG. So let that the causal graph G_P be given by the DAG in Figure 1. By the Markov factorization (1), we know that $P(\mathbf{X}) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1)$.*

Now assume that we do not have access to this joint distribution, but instead we are only provided with two marginal distributions $P(X_1, X_2), P(X_1, X_3)$ and the knowledge of the underlying DAG. Even with no interventions, observing the two marginal distributions is sufficient to recover the joint distribution $P(X_1, X_2, X_3)$, since it is possible to separately estimate $P(X_2 \mid X_1)$, $P(X_3 \mid X_1)$ from the marginals and $P(X_1)$ from either of the marginals. This phenomenon is correctly captured in our IKL metric by modelling unobserved variables as interventions:

$$D_{\mathrm{KL}}(P(X_1, X_2) \mid Q(X_1, X_2)) + D_{\mathrm{KL}}(P(X_1, X_3) \mid Q(X_1, X_3)) \tag{7}$$

$$\begin{aligned} &= D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) + \mathbb{E}_{P(X_1)}[D_{\mathrm{KL}}(P(X_2 \mid X_1) \| Q(X_2 \mid X_1)) \\ &\quad + D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) + \mathbb{E}_{P(X_1)}[D_{\mathrm{KL}}(P(X_3 \mid X_1) \| Q(X_3 \mid X_1)) \end{aligned} \tag{8}$$

2. Specifically, they need to coincide on the union of support sets $\mathrm{supp}(P(\mathbf{PA}_i^{G_P})) \cup \mathrm{supp}(P^e(\mathbf{PA}_i^{G_P}))$

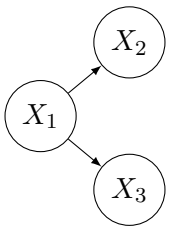

Figure 1: A sample DAG over variables X_1, X_2, X_3.

If an intervention happened over the unobserved variables, say $P^{e_1}(\mathbf{X}) = P(X_1, X_2)\tilde{P}(X_3)$, $P^{e_2}(\mathbf{X}) = P(X_1, X_3)\tilde{P}(X_2)$, *we obtain:*

$$= 2 \cdot D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1,e_2\}} \| \mathscr{Q}) \tag{9}$$

We can now minimize each of the terms in equation (8) with respect to our model distribution Q. *By Remark 10, this will also minimize the distance between the joint distributions* P, Q *without observing all variables at once. Note that the term* $D_{\mathrm{KL}}(P(X_1) \| Q(X_1))$ *occurs twice in equation (8). This means that we could still identify the joint distribution in the same way, if an intervention had happened on* X_1 *in either* $P(X_1, X_2)$ *or* $P(X_1, X_3)$.

3.3. Interventional Differences in the IKL divergence

In the previous section, we discussed how the Interventional KL divergence quantifies distributional differences between two causal models $\mathscr{P} = (P, G_P)$ and $\mathscr{Q} = (Q, G_Q)$. Distributional differences fully characterize the distance between $\mathscr{P}$ and $\mathscr{Q}$ if the causal graphs are identical, i.e. $G_P = G_Q$. Since this is generally not the case, we now study how structural differences, i.e. $G_P \neq G_Q$ show up in the IKL divergence. In the following, we assume that the empty intervention $\mathcal{I} = \emptyset$ is included in the environment.

Our current definition of the IKL divergence depends on the decomposition of the KL divergence in Lemma 8. However, this requires access to the true causal graph G_P, which determines the parent sets that we condition on. We now generalize this lemma to using a (potentially different) factorization determined by the model graph G_Q. It turns out that this decomposition still holds as long as P is Markovian with respect to G_Q. Otherwise, there is a residual term quantifying the difference of P from being Markovian, for which we introduce the notion of a *Markov Projection*.

Definition 12 (Markov Projection) *Let* $\mathcal{A}$ *be the space of probability distributions over the variable set* $\mathbf{X}$. *The* Markov projection *(or* M-projection*) of a probability distribution* P *onto a DAG* G *is the mapping* $\pi_G : \mathcal{A} \to \mathcal{A}$ *defined by:*

$$\pi_G(P) = \underset{Q \text{ Markov w.r.t. } G}{\arg\min} D_{\mathrm{KL}}(P \| Q)$$

The following Lemma 13 explicitly describes the Markov projection onto a DAG G.

Lemma 13 *Given a distribution* P *and a DAG* G*:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{PA}_i^G). \tag{10}$$

From the Markov factorization property 2 it follows that $P = \pi_G(P)$ if and only if P is Markovian with respect to G. Otherwise, we have $P \neq \pi_G(P)$, since there are variables X_i, X_j such that $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{PA}_i^G$ with respect to P, but $X_i \perp\!\!\!\perp X_j \mid \mathbf{PA}_i^G$ in $\pi_G(P)$. If we now allow for general (non-Markovian) distributions P in Lemma 8, the deviation from being Markovian enters the equation as an additional term:

Lemma 14 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs and assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = &\sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) \end{aligned} \tag{11}$$

As a special case we have for $\mathcal{I}_e = \emptyset$

$$\begin{aligned} D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = &\sum_{i \in [d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ D_{\mathrm{KL}}(P(\mathbf{X}) \| \pi_{G_Q}(P)(\mathbf{X})) \end{aligned} \tag{12}$$

Let us consider two cases in which G_P and G_Q may be structurally different: If G_P induces a conditional independence relationship that is not entailed by G_Q, by faithfulness 3, it follows that P is necessarily different from Q. This difference then shows up in the D_{IKL} divergence via a non-zero term $D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q})) > 0$. If otherwise, there is a conditional independence relationship induced by G_Q, which not satisfied by G_P, P will not be Markovian with respect to G_Q and the residual term $D_{\mathrm{KL}}(P(\mathbf{X}) \| \pi_{G_Q}(P)(\mathbf{X})) > 0$. We now adapt the previous definition 9 of the Interventional KL divergence correspondingly:

Definition 15 (Interventional KL divergence) *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be two general causal models. Further, assume that we have access to a set of interventional distributions $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$ generated from $\mathscr{P}$. We define the Interventional KL divergence as*

$$\begin{aligned} D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \Bigg[&\sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) \Bigg] \end{aligned} \tag{13}$$

Note that this definition is consistent with our previous notion of interventional distance in the case where $G_P = G_Q$, see Definition 9.

The definition of the D_{IKL} went under the premise that the effect of the interventions that we consider is unknown. i.e. only the joint distributions $P^e(\mathbf{X})$ and intervention targets $\mathcal{I}_e$ are observed, but we don't know the mapping $P(X_i \mid \mathbf{PA}_i^{G_P}) \mapsto \tilde{P}(X_i \mid \mathbf{PA}_i^{G_P})$. If we know the mechanism shifts that happened between the distribution $P(\mathbf{X})$ and the interventional distributions $P^e(\mathbf{X})$ and we are able to reproduce them in our model $\mathscr{Q} = (Q, G_Q)$, then the D_{IKL} collapses to a simpler form, namely the average of the KL divergences between the joint distributions $P^e(\mathbf{X})$ and $Q(\mathbf{X})$.

Theorem 16 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ with P Markovian w.r.t G_Q and a set of interventional distributions $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i \notin \mathcal{I}_e} Q(X_i \mid \mathbf{PA}_i^{G_Q}) \cdot \prod_{j \in \mathcal{I}_e} P^e(X_j \mid \mathbf{PA}_j^{G_P}) \tag{14}$$

where $P^e(X_j \mid \mathbf{PA}_j^{G_P})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{15}$$

Proof [Sketch] If we apply the same mechanism shifts to Q^e that appear between P and P^e, we can use the decomposition of the KL divergence in Lemma 8 to see that these terms cancel out. The sum over the remaining terms then equals our definition of the Interventional KL divergence. ■

In particular, this Theorem applies to hard interventions on single variables that were considered by Peyrard and West (2020).

3.4. Interventional differences beyond Markov equivalence

With our analysis so far, we have discussed how distributional and statistically identifiable differences in the causal graphs are encoded in the IKL divergence. For the remainder of this section, we demonstrate how structural differences beyond Markov equivalence show up in the Interventional KL divergence. We make use of our assumptions on the multi-environment distributions (Assumptions 6, 7), which imply for Markov equivalent G_P, G_Q that

$$\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})} D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| P(X_i \mid \mathbf{PA}_i^{G_Q})) = 0 \quad \text{for any } e \text{ with } i \notin \mathcal{I}_e$$
$$\text{if and only if } \mathbf{PA}_i^{G_Q} = \mathbf{PA}_i^{G_P}$$

As shown in Example 21, this equivalence does not hold given only a limited set of interventional distributions. A comprehensive characterization of differences between $P^e(X_i \mid \mathbf{PA}_i^{G_Q})$ and $P(X_i \mid \mathbf{PA}_i^{G_Q})$ can be given based on d-separation in the augmented DAG $G_{P,\mathbf{X} \cup \{e\}}$ (Perry et al., 2022). Since we assumed G_P to be generally unknown, we here propose a simpler sufficient condition that trades off knowledge about G_P with access to interventional distributions. It is based on the following definition of directed paths between a variable X_i and intervened variables.

Definition 17 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs and P^e be an interventional distribution generated from $\mathscr{P}$. For some (potentially different) DAG G, we say that there is a directed path $e \overset{G}{\to} \mathbf{PA}_i^{G_Q}$ in G if there exists $j \in \mathcal{I}_e$ and a transitive edge $X_j \to \ldots \to C$ in G for some $C \in \mathbf{PA}_i^{G_Q}$.*

So, $e \overset{G_P}{\to} \mathbf{PA}_i^{G_Q}$ means that there is a downstream connection between an intervened variable and (at least) one variable in $\mathbf{PA}_i^{G_Q}$. Since it depends on the graph whether such a downstream path between two variables exists, we specify the graph above the arrow. If G_Q shares the same skeleton with G_P, this definition yields a sufficient condition for the mechanism $P^e(X_i \mid \mathbf{PA}_i^{G_Q})$ to be different from $P(X_i \mid \mathbf{PA}_i^{G_Q})$.

Lemma 18 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton. Further, let P^e be an interventional distribution generated from $\mathscr{P}$. If there exists a conditioned child $C \in \mathbf{PA}_i^{G_Q} \cap \mathbf{CH}_i^{G_P}$ and a directed path $e \overset{G_P}{\to} C$, then we have that $P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \neq P(X_i \mid \mathbf{PA}_i^{G_Q})$.*

Assume that one of the parents $C \in \mathbf{PA}_i^{G_Q}$ of X_i in G_Q truly was a child of X_i in G_P, i.e. the arrow $C \to X_i$ was mis-oriented between G_Q and G_P. According to this Lemma, if there is a directed path $e \overset{G_P}{\to} C$, then the corresponding term in the Interventional KL divergence will be positive. Since this a condition on the graphical structure of G_P, it cannot be verified if G_P is unknown. But as it turns out, it is sufficient if such a path exists in G_Q provided none of the intermediate variables are intervened. We formalize these observations in the following Theorem:

Theorem 19 *Let $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$ be a collection of distributions generated from a CGM (P, G_P) and a model CGM (Q, G_Q) be given. Let further the set of variables with mis-oriented edges be given by*

$$\mathcal{B} = \{j \in [d] \mid X_i \to X_j \text{ in } G_Q, X_j \to X_i \text{ in } G_P\} \tag{16}$$

Now assume that the following condition holds: For some $j \in \mathcal{B}$ there exists an environment $e \in \mathcal{E}$ with a directed path $e \overset{G_Q}{\to} \mathbf{PA}_j^{G_Q}$ in G_Q such that only the first variable on the path is intervened in e. Then we have that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ if and only if $(P, G_P) = (Q, G_Q)$.

Proof [sketch] If $\mathscr{P} = \mathscr{Q}$, we can deduce $D_{\mathrm{IKL}}(\mathscr{P} \| \mathscr{Q}) = 0$ from our Assumptions on the environment 6 and pseudo-causal faithfulness 7. Conversely, if $P \neq Q$, $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \geq 1/|\mathcal{E}| \; D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. This case then follows from Lemma 18, since a directed path to the parents of a mis-oriented edge in G_Q gives rise to a directed path to a mis-oriented edge in G_P.

■

As a consequence of Theorem 19, we can deduce a sufficient condition on the intervention targets that only depends on our model estimate (Q, G_Q) to provably quantify any observational or interventional differences with respect to subsets of all observed variables.

Corollary 20 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathbf{X}_S \subseteq \mathbf{X}$ be a subset of variables for some index set $S \subseteq [d]$. Assume that for all $i \in S$ there exists an environment $e \in \mathcal{E}$ with a directed path $e \overset{G_Q}{\to} \mathbf{PA}_i^{G_Q}$ such that only the first variable on the path is intervened upon. Then, the restricted IKL divergence summing over the sets $\mathbf{X}_S \cap \mathcal{I}_e^C$ is zero[3] if and only if $\mathscr{P}\big|_{\mathbf{X_S}} = \mathscr{Q}\big|_{\mathbf{X_S}}$. This means that the subgraphs over variables $\bigcup_{i \in S} \{X_i, \mathbf{PA}_i^{G_Q}\}$ are the same and the corresponding mechanisms $P(X_i \mid \mathbf{PA}_i^{G_Q}), Q(X_i \mid \mathbf{PA}_i^{G_Q})$ coincide.*

Example 21 *Let $(P, G_P), (Q, G_Q)$ be two CGMs with $P = Q$ and graphs given in Figure 2. We want to compute the interventional KL divergence to evaluate the fit of (Q, G_Q) to (P, G_Q) using interventional distributions P^{e_1}, P^{e_2}.*

3. $\mathcal{I}_e^C$ denotes the complement set of $\mathcal{I}_e$

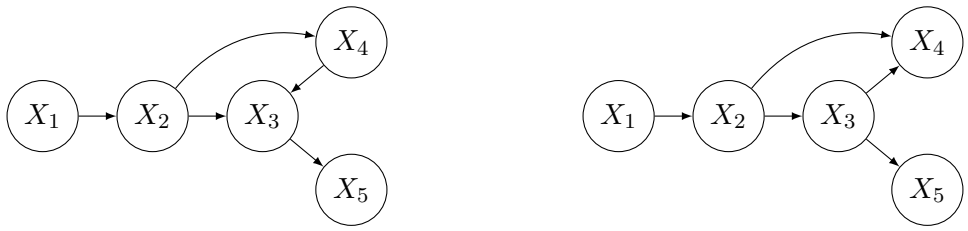

Figure 2: Two Markov equivalent DAGs G_P (left) and G_Q (right) over variables $X_1, \ldots, X_5$.

Let P^{e_1} be given by a single mechanism shift $P(X_5|X_3) \mapsto \tilde{P}(X_5|X_3)$. Further, assume that P^{e_2} emerges from a single mechanism shift $P(X_2|X_1) \mapsto \tilde{P}(X_2|X_1)$. Computing the Interventional KL divergence now yields:

$$\begin{aligned} D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1,e_2\}} \| \mathscr{Q}) &= \frac{1}{2}[0 + \mathbb{E}_{P^{e_2}(X_2,X_3)} D_{\mathrm{KL}}(P^{e_2}(X_4|X_2,X_3) \| Q(X_4|X_2,X_3))] \\ &> 0 \end{aligned}$$

We see that the contribution from the first interventional distribution P^{e_1} vanishes, since all the conditional distributions remain invariant between P and P^{e_1}. However, by Corollary 20 the second intervention should be sufficient for telling the causal models apart. Indeed, we have a non-zero term $\mathbb{E}_{P^{e_2}(X_2,X_3)} D_{\mathrm{KL}}(P^{e_2}(X_4|X_2,X_3) \| Q(X_4|X_2,X_3))] > 0$ by Lemma 18.

4. Discussion

Multi-environment distributions provide a natural setting for machine learning algorithms to both learn causal structures and leverage them to achieve robustness and out-of-distribution generalization. We have assayed how knowledge about the true causal graph can be used to facilitate learning under distribution shifts and generalize from marginal observations to properties of the joint distribution. Both of these are significant problems for learning in real world settings, where we cannot always afford to collect a large dataset suited to a task. Conversely, multi-domain data can provide a useful learning signal to drive causal discovery.

The focus of the current paper has been a theoretical exploration of properties of the Interventional KL divergence, and applications are thus beyond our scope. Nevertheless, we would like to discuss some possible use cases to point out directions for future work.

Interventional KL Divergence and Estimation The following result captures the operational significance of low IKL divergence.

Theorem 22 *Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ with P Markovian w.r.t. G_Q and a set of interventional distributions $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as in Theorem 16.*

Suppose $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \leq \epsilon$. Then, for any bounded function f mapping to $[-B, B]$ and any parameter $\rho > 0$, for at least $1 - \rho$ fraction of the environments $e \in \mathcal{E}$:

$$\left| \mathbb{E}_{P^e(\mathbf{X})}[f(\mathbf{X})] - \mathbb{E}_{Q^e(\mathbf{X})}[f(\mathbf{X})] \right| \leq \frac{B\sqrt{\epsilon}}{\rho}$$

In other words, if the IKL divergence is small, then for most environments, executing the same mechanism changes in 'Nature' and in our model $\mathscr{Q}$ would yield similar statistics. Thus, the more heterogeneous the set of environments, the more similar $\mathscr{P}$ and $\mathscr{Q}$ are in terms of their causal implications.

Evaluation tool for causal discovery methods If data from multi-environment distributions are available, there is a range of methods to recover the true causal structure asymptotically (Perry et al., 2022; Janzing et al., 2012; Huang et al., 2019; He and Geng, 2016; Rojas-Carulla et al., 2015; Arjovsky et al., 2019; Krueger et al., 2021). Here, the IKL divergence could help develop methods to verify a learned causal model, provided that intervention targets are known for some of the environments. This can then take into account both the interventional differences, as encoded by the learned DAG, and the observational differences encoded in the learned conditional distributions.

Active Learning of Causal Structures If we do not have a sufficient number of environment distributions a priori to recover the true causal graph, but intervention targets are more broadly available, we have shown in Corollary 20 that one could evaluate the fit of a model to the true causal model with respect to subsets of variables. These partially learned and verified models could then be used for inference until more interventional data become available to uncover further causal relationships. Indeed, our metric can even serve for orienting edges in the graph: If we compute $D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1\}} \| (P, G_Q))$, then any non-zero term in the IKL divergence is necessarily caused by mis-oriented edges. Specifically, if for some variable $i \in [d] \backslash \mathcal{I}_e$ we have

$$\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})} D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| P(X_i \mid \mathbf{PA}_i^{G_Q})) > 0,$$

then $\mathbf{PA}_i^{G_Q} \neq \mathbf{PA}_i^{G_P}$. Thus, we can identify the true parent set by re-computing this term with respect to parent sets $\mathbf{PA}_i^{G}$ for all graphs G in the Markov equivalence class of G_Q.

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions, 2018. URL https://arxiv.org/abs/1805.09697.

S. Acid and L. M. De Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, may 2003. doi: 10.1613/jair.1061. URL https://doi.org/10.1613%2Fjair.1061.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint 1907.02893*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization, 2022. URL https://arxiv.org/abs/2207.09944.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 107–114, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL https://proceedings.mlr.press/v2/eaton07a.html.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6ad4174eba19ecb5fed17411a34ff5e6-Paper.pdf.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. *arXiv e-prints*, art. arXiv:2203.15756, March 2022.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. 2011. doi: 10.48550/ARXIV.1104.2808. URL https://arxiv.org/abs/1104.2808.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 77(1):291–318, 2015. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/24774735.

Yango He and Zhi Geng. Causal network learning from multiple interventions of unknown manipulated targets, 2016. URL https://arxiv.org/abs/1610.08611.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data with independent changes. 2019. doi: 10.48550/ARXIV.1903.01672. URL https://arxiv.org/abs/1903.01672.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.01.002. URL https://www.sciencedirect.com/science/article/pii/S0004370212000045.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2010.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. 2016. doi: 10.48550/ARXIV.1611.10351. URL https://arxiv.org/abs/1611.10351.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis, 2022. URL https://arxiv.org/abs/2206.02013.

Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. 2013. doi: 10.48550/ARXIV.1306.1043. URL https://arxiv.org/abs/1306.1043.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Maxime Peyrard and Robert West. A ladder of causal distances, 2020. URL https://arxiv.org/abs/2005.02480.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. 2015. doi: 10.48550/ARXIV.1507.05333. URL https://arxiv.org/abs/1507.05333.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning, 2012. URL https://arxiv.org/abs/1206.6471.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81. 01 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2021. URL https://arxiv.org/abs/2106.04619.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–819–III–827. JMLR.org, 2013.

Appendix A. Full proofs

A.1. Proof of Lemmas 8 and 14

Lemma 8 *Let P, Q be distributions sharing the same causal graph G_P and assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = &\sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P}))] \\ &+ \underbrace{\sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P}))]}_{=:D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})} \end{aligned} \tag{4}$$

As a special case we have for $\mathcal{I}_e = \emptyset$:

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i\in[d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P}))] \tag{5}$$

Lemma 14 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs and assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = &\sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ \sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ D_{\mathrm{KL}}(P^e(\mathbf{X})\|\pi_{G_Q}(P^e)(\mathbf{X})) \end{aligned} \tag{11}$$

As a special case we have for $\mathcal{I}_e = \emptyset$

$$\begin{aligned} D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = &\sum_{i\in[d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ &+ D_{\mathrm{KL}}(P(\mathbf{X})\|\pi_{G_Q}(P)(\mathbf{X})) \end{aligned} \tag{12}$$

Proof We show the general decomposition of $D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X}))$ in equation 12 using the fact that Q is Markovian with respect to G_Q. The decomposition for $D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X}))$ then just follows by splitting up the sums. Also note that Lemma 8 follows as a special case.

$$\begin{aligned}
D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) &= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{P\pi_G(P)}{Q\pi_G(P)}\right] \\
&= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{\pi_G(P)}{Q}\right] + \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{P}{\pi_G(P)}\right] \\
&= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{\pi_G(P)}{Q}\right] + D_{\mathrm{KL}}(P(\mathbf{X})\|\pi_G(P)(\mathbf{X}))
\end{aligned}$$

We now proceed to analyze the remaining term using the characterization of the Markov projection in Lemma 13.

$$\begin{aligned}
\mathbb{E}_{P(\mathbf{x})}\left[\log \frac{\pi_G(P)}{Q}\right] &= \mathbb{E}_{P(\mathbf{x})}\left[\log \prod_{i=1}^{d} \frac{\pi_G(P)(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)}\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{x})}\left[\log \frac{P(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)}\right] \\
&= \sum_{i=1}^{d} \sum_{X_i, \mathbf{PA}_i} \log \frac{P(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)} \\
&\quad \cdot \sum_{\mathbf{X}_{[d]} \setminus \{X_i, \mathbf{PA}_i\}} P(X_i, \mathbf{PA}_i) \cdot P(\mathbf{X}_{[d]} \setminus \{X_i, \mathbf{PA}_i\} \mid X_i, \mathbf{PA}_i) \\
&= \sum_{i=1}^{d} \sum_{X_i, \mathbf{PA}_i} \log \frac{P(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)} \cdot P(X_i \mid \mathbf{PA}_i) \cdot P(\mathbf{pa}_i) \\
&\quad \cdot \underbrace{\sum_{\mathbf{X}_{[d]} \setminus \{X_i, \mathbf{PA}_i\}} P(\mathbf{X}_{[d]} \setminus \{X_i, \mathbf{PA}_i\} \mid X_i, \mathbf{PA}_i)}_{=1 \text{ for every choice of } (X_i, \mathbf{PA}_i)} \\
&= \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{PA}_i^G)\|Q(X_i \mid \mathbf{PA}_i^G))]
\end{aligned}$$

To simplify notation, we here leave away the supscript G_Q for the parents and use sums for the expected values, but the continuous case follows with integrals analogously.

■

A.2. Proof of Lemma 13

Lemma 13 *Given a distribution P and a DAG G:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{PA}_i^G). \tag{10}$$

Proof Given distribution P and DAG G, we have

$$\begin{aligned}
\pi_G(P)(\mathbf{X}) &= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \; D_{\mathrm{KL}}(P(\mathbf{X}) \| \tilde{P}(\mathbf{X})) \\
&= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^G) \| \tilde{P}(X_i \mid \mathbf{pa}_i^G))] + D_{\mathrm{KL}}(P(\mathbf{X}) \| \pi_G(P)(\mathbf{X})) \\
&= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)}[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^G) \| \tilde{P}(X_i \mid \mathbf{pa}_i^G))]
\end{aligned}$$

where we use Lemma 14 in line 2. The claim can then be deduced from noting that $\tilde{P}(\mathbf{X}) = \prod_{i=1}^{d} P(X_i | \mathbf{PA}_i^G)$ sets the sum to zero. ■

A.3. Proof of Theorem 16

Theorem 16 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ with P Markovian w.r.t G_Q and a set of interventional distributions $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i \notin \mathcal{I}_e} Q(X_i \mid \mathbf{PA}_i^{G_Q}) \cdot \prod_{j \in \mathcal{I}_e} P^e(X_j \mid \mathbf{PA}_j^{G_P}) \tag{14}$$

where $P^e(X_j \mid \mathbf{PA}_j^{G_P})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{15}$$

Proof Since P is Markovian w.r.t. G_Q, so is any interventional distribution P^e w.r.t. G_{Q_e}, where G_{Q_e} is the post-intervention graph after possibly deleting edges (e.g. as the result of hard-interventions) with respect to G_Q. We can decompose the RHS using Lemma 8. This gives us

$$\begin{aligned}
&\frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \\
&= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\sum_{i=1}^{d} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_{Q_e}})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_{Q_e}}) \| Q^e(X_i \mid \mathbf{pa}_i^{G_{Q_e}}))]] \right]
\end{aligned}$$

By assumption, we now have that $\mathbf{PA}_i^{G_{Q_e}} = \mathbf{PA}_i^{G_Q}$ for all $i \in [d] \backslash \mathcal{I}_e$, and $\mathbf{PA}_i^{G_{Q_e}} = \mathbf{PA}_i^{G_P}$ otherwise. It follows, that

$$Q^e(X_i \mid \mathbf{PA}_i^{G_{Q_e}}) = \begin{cases} P^e(X_i \mid \mathbf{PA}_i^{G_P}), & \text{if } i \in \mathcal{I}_e \\ Q(X_i \mid \mathbf{PA}_i^{G_Q}), & \text{if } i \notin \mathcal{I}_e \end{cases}$$

Thus, term above simplifies to

$$\begin{aligned}
&= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\sum_{i \in [d] \backslash \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))]] \right] \\
&= D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})
\end{aligned}$$

■

A.4. Proof of Lemma 18

Lemma 18 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton. Further, let P^e be an interventional distribution generated from $\mathscr{P}$. If there exists a conditioned child $C \in \mathbf{PA}_i^{G_Q} \cap \mathbf{CH}_i^{G_P}$ and a directed path $e \overset{G_P}{\to} C$, then we have that $P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \neq P(X_i \mid \mathbf{PA}_i^{G_Q})$.*

Proof The proof is based on a Corollary 4.4. by Perry et al. (2022). Since there exists a conditioned child $C \in \mathbf{PA}_i^{G_Q} \cap \mathbf{CH}_i^{G_P}$ and a directed path $e \overset{G_p}{\to} C$, case 3 applies. Thus, it follows that $P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \neq P(X_i \mid \mathbf{PA}_i^{G_Q})$. ■

A.5. Proof of Theorem 19

Theorem 19 *Let $\mathcal{P} = (P^e)_{e \in \mathcal{E}}$ be a collection of distributions generated from a CGM (P, G_P) and a model CGM (Q, G_Q) be given. Let further the set of variables with mis-oriented edges be given by*

$$\mathcal{B} = \{j \in [d] \mid X_i \to X_j \text{ in } G_Q, X_j \to X_i \text{ in } G_P\} \tag{16}$$

Now assume that the following condition holds: For some $j \in \mathcal{B}$ there exists an environment $e \in \mathcal{E}$ with a directed path $e \overset{G_Q}{\to} \mathbf{PA}_j^{G_Q}$ in G_Q such that only the first variable on the path is intervened in e. Then we have that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ if and only if $(P, G_P) = (Q, G_Q)$.

Proof Let's first assume that $(P, G_P) = (Q, G_Q)$. By our assumption on the multi-environment distributions 6 and 7, we have

$$P^e(X_i \mid \mathbf{PA}_i^{G_Q}) = P^e(X_i \mid \mathbf{PA}_i^{G_P}) = P(X_i \mid \mathbf{PA}_i^{G_P}) = Q(X_i | \mathbf{PA}_i^{G_P}) \tag{17}$$

if and only if $i \in [d] \backslash \mathcal{I}_e$. Thus, all the conditional divergences in the IKL divergence vanish. Since $G_P = G_Q$, P is Markovian w.r.t. G_P and so are all interventional distributions, which even factorize over subgraphs of G_P. Thus, all the residual terms $KL(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) = 0$ and therefore $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$.

Conversely, assume that $(P, G_P) \neq (Q, G_Q)$. If $P \neq Q$, we have that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \geq 1/|\mathcal{E}| D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. If $P = Q$, then G_Q is necessarily Markov equivalent to G_P and thus G_Q, G_P share the same skeleton and v-structures. Since $G_P \neq G_Q$, there exists a variable X_i such that $X_j \to X_i$ in G_P, but $X_i \to X_j$ in G_Q. By assumption, we know there exists an environment e such that there is a directed path $e \overset{G_Q}{\to} \mathbf{PA}_j^{G_Q}$ in G_Q, i.e. there is a parent $C \in \mathbf{PA}_j^{G_Q}$ and a path $\hat{X} \to \ldots \to C$ in G_Q such that $\hat{X}$ is intervened. W.l.o.g. we can assume that this also yields a directed path in G_P: If it did not, i.e. there one of the edges was oriented in the wrong way, say $\hat{X} \overset{G_Q}{\to}, \ldots, \overset{G_Q}{\to}, X_k \overset{G_Q}{\to} X_l$, but $\hat{X} \overset{G_P}{\to}, \ldots, \overset{G_P}{\to}, X_k \overset{G_P}{\leftarrow} X_l$, then $l \in \mathcal{B}$ and by assumption $l \notin \mathcal{I}_e$ and the argument below applies to this edge.

If $C = X_i$, i.e. the variable $C \in \mathbf{PA}_j^{G_Q}$ indeed leads to a mis-oriented edge in G_P, Lemma 18 applies immediately. If not and $C \to X_j$ is correct in both G_P and G_Q, and $\hat{X} \to C \to X_j \to X_i$ gives us again a directed path to the conditioned child X_i. Summarizing, by Lemma 18 we have that $P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \neq P(X_j \mid \mathbf{PA}_j^{G_Q}) = Q(X_j \mid \mathbf{PA}_j^{G_Q})$,. Finally, this gives us

$$\mathbb{E}_{P(\mathbf{PA}_j^{G_Q})} D_{\mathrm{KL}}(P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \| Q(X_j \mid \mathbf{PA}_j^{G_Q})) > 0$$

so that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$, since $j \notin \mathcal{I}_e$. ■

A.6. Proof of Corollary 20

Corollary 20 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathbf{X}_S \subseteq \mathbf{X}$ be a subset of variables for some index set $S \subseteq [d]$. Assume that for all $i \in S$ there exists an environment $e \in \mathcal{E}$ with a directed path $e \overset{G_Q}{\to} \mathbf{PA}_i^{G_Q}$ such that only the first variable on the path is intervened upon. Then, the restricted IKL divergence summing over the sets $\mathbf{X}_S \cap \mathcal{I}_e^C$ is zero[4] if and only if $\mathscr{P}\big|_{\mathbf{X_S}} = \mathscr{Q}\big|_{\mathbf{X_S}}$. This means that the subgraphs over variables $\bigcup_{i \in S}\{X_i, \mathbf{PA}_i^{G_Q}\}$ are the same and the corresponding mechanisms $P(X_i \mid \mathbf{PA}_i^{G_Q}), Q(X_i \mid \mathbf{PA}_i^{G_Q})$ coincide.*

Proof Since $P = Q$, we have that G_P, G_Q are Markov equivalent, so they share the same skeleton. If $\mathscr{P}\big|_{\mathbf{X_S}} = \mathscr{Q}\big|_{\mathbf{X_S}}$, then it holds that $\mathbf{PA}_i^{G_Q} = \mathbf{PA}_i^{G_P}$ for all $i \in S$. For all e with $i \in \mathbf{X}_S \cap \mathcal{I}_e^C$, we then have that $P^e(X_i|\mathbf{PA}_i^{G_Q}) = P(X_i|\mathbf{PA}_i^{G_Q}) = Q(X_i|\mathbf{PA}_i^{G_Q})$ and the restricted IKL divergence

$$\begin{aligned} D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \Bigg[\sum_{i \in \mathbf{X}_S \cap \mathcal{I}_e^C} & \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}[D_{\mathrm{KL}}(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}))] \\ & + D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) \Bigg] = 0 \end{aligned} \tag{18}$$

vanishes. Conversely, let $\mathscr{P}\big|_{\mathbf{X_S}} \neq \mathscr{Q}\big|_{\mathbf{X_S}}$. Thus, there exists a mis-oriented edge in the subgraphs as defined above. By our assumption, the condition in Theorem 19 is satisfied for all $i \in S$. Thus, it is in particular satisfied for the mis-oriented edge. By the proof of Theorem 19, we can conclude that this leads to a non-zero term in the IKL divergence for the corresponding environment and the variable $X_j \in \mathbf{X}_S$, thus also implying that the restricted KL divergence is greater than zero.

■

A.7. Proof of Theorem 22

Proof By Theorem 16:

$$\frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \leq \epsilon$$

4. $\mathcal{I}_e^C$ denotes the complement set of $\mathcal{I}_e$

Using Pinsker's inequality and the concavity of the square-root function:

$$\begin{aligned}\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) &\leq \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sqrt{D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))/2} \\ &\leq \sqrt{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))} \\ &\leq \sqrt{\epsilon}\end{aligned}$$

By Markov's inequality, for at least $1-\rho$ fraction of e's,

$$D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) \leq \frac{\sqrt{\epsilon}}{\rho}.$$

The claim now follows from the boundedness of f. ■